UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

MODELO HÍBRIDO PARA LA DESCRIPCIÓN DE ESCENAS USANDO APRENDIZAJE PROFUNDO

TESIS PARA OBTENER EL GRADO DE:

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

MARCO ANTONIO LÓPEZ SÁNCHEZ

BAJO LA DIRECCIÓN DE:

DR. OSCAR ALBERTO CHÁVEZ BOSQUEZ

EN CODIRECCIÓN:

DR. JOSÉ HERNÁNDEZ TORRUCO

CUNDUACÁN, TABASCO, A: DICIEMBRE 2025

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

MODELO HÍBRIDO PARA LA DESCRIPCIÓN DE ESCENAS USANDO APRENDIZAJE PROFUNDO

TESIS PARA OBTENER EL GRADO DE:

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

MARCO ANTONIO LÓPEZ SÁNCHEZ

BAJO LA DIRECCIÓN DE:

DR. OSCAR ALBERTO CHÁVEZ BOSQUEZ

EN CODIRECCIÓN:

DR. JOSÉ HERNÁNDEZ TORRUCO

CUNDUACÁN, TABASCO, A: DICIEMBRE 2025

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán el día Diez del mes de Diciembre del año 2025, el que suscribe **Marco Antonio López Sánchez**, alumno del Programa de la **Doctorado en Ciencias de la Computación** con número de matrícula **201H18003**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención de Grado de Doctorado y titulada **Modelo híbrido para la descripción de escenas usando Aprendizaje Profundo**, dirigida por el Dr. Oscar Alberto Chávez Bosquez y el Dr. José Hernández Torruco.

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco a 10 de Diciembre de 2025.

Estudiante: Marco Antonio López Sánchez

Cunduacán, Tabasco, a 08 de enero de 2026
Oficio No. 0044/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Marco Antonio López Sánchez
Egresado del Doctorado en Ciencias de la Computación

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional ***"Modelo hibrido para la descripción de escenas usando aprendizaje profundo"***, para presentar examen y obtener el Grado de Doctor en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

Dr. Óscar Alberto González González
Director

C.c.p. Mtra. Yenny Lorena Dussán Rojas. – Encargada del despacho de la Coordinación de Posgrado.
Archivo.
Consecutivo.

DR.*OAGG/YLDR

Carta de Cesión de Derechos

Villahermosa, Tabasco a 10 de Diciembre de 2025.

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Modelo híbrido para la descripción de escenas usando Aprendizaje Profundo**.

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención de que tenemos el derecho al reconocimiento como autores de la obra, y a la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR

Estudiante: Marco Antonio López Sánchez

TESTIGOS

Dr. Oscar Alberto Chávez Bosquez

Dr. José Hernández Torruco

Índice general

Índice de figuras

Índice de tablas

Resumen

La descripción automática de escenas es una tarea compleja que requiere la integración de técnicas de visión por computadora y procesamiento de lenguaje natural. Esta tesis doctoral propone un modelo híbrido basado en arquitecturas *encoder-decoder*, combinando redes neuronales convolucionales (CNN) para la extracción de características visuales y redes LSTM para la generación secuencial de descripciones en lenguaje natural. El trabajo se estructura en torno a tres contribuciones principales: (i) un estudio comparativo sobre el impacto de diferentes algoritmos de optimización (SGD, RMSprop y Adam) en el entrenamiento de CNN para la clasificación binaria de imágenes, (ii) una revisión sistemática de la literatura sobre arquitecturas *encoder-decoder* aplicadas a la descripción automática de imágenes, abarcando 53 artículos publicados entre 2014 y 2022, y (iii) el diseño y evaluación de un sistema de reconocimiento facial basado en aprendizaje profundo, validado en condiciones del mundo real.

Los resultados muestran que el optimizador Adam supera a otros algoritmos en tareas de clasificación, que la combinación CNN+LSTM sigue siendo la arquitectura predominante en tareas de *captioning*, y que los modelos propuestos son robustos bajo condiciones adversas. Se discute la relevancia de métricas de evaluación como BLEU, METEOR y CIDEr, así como la necesidad de avanzar hacia evaluaciones más fundamentadas semánticamente e interpretables. Finalmente, se proponen direcciones futuras de investigación, que incluyen la exploración de modelos basados en *transformers*, la reducción de la dependencia de datos etiquetados y la mejora de la explicabilidad en sistemas generativos. Esta tesis proporciona fundamentos teóricos y empíricos para el desarrollo de sistemas multimodales más eficientes, interpretables y aplicables en entornos reales.

Palabras clave: Descripción Automática de Imágenes, Aprendizaje Profundo, Redes Neuronales Convolucionales.

Abstract

Automatic scene description is a complex task that requires the integration of computer vision techniques and natural language processing. This doctoral dissertation proposes a hybrid model based on encoder-decoder architectures, combining Convolutional Neural Networks (CNN) for visual feature extraction and LSTM networks for the sequential generation of natural language descriptions. The work is structured around three main contributions: (i) a comparative study on the impact of different optimization algorithms (SGD, RMSprop, and Adam) in training CNNs for binary image classification, (ii) a systematic literature review of encoder-decoder architectures applied to image captioning, covering 53 articles published between 2014 and 2022, and (iii) the design and evaluation of a deep learning-based facial recognition system validated under real-world conditions.

The results show that the Adam optimizer outperforms other algorithms in classification tasks, that CNN+LSTM remains the predominant architecture in captioning tasks, and that the proposed models are robust under adverse conditions. The relevance of evaluation metrics such as BLEU, METEOR, and CIDEr is discussed, as well as the need to move toward more semantically grounded and interpretable assessments. Finally, future research directions are proposed, including the exploration of transformer-based models, the reduction of labeled data dependency, and the enhancement of explainability in generative systems. This thesis provides theoretical and empirical foundations for the development of more efficient, interpretable, and applicable multimodal systems in real-world environments.

Keywords: Automatic Image Description, Deep Learning, Convolutional Neural Networks

Capítulo 1

Generalidades

1.1. Introducción

La comprensión automática del contenido visual constituye uno de los desafíos fundamentales en el campo de la inteligencia artificial, particularmente en la intersección entre la visión por computadora y el procesamiento del lenguaje natural (Vinyals et al., 2015; Bernardi et al., 2016). La descripción automática de imágenes —también conocida como *image captioning*— es una tarea compleja que busca generar, en lenguaje natural, oraciones que representen con fidelidad el contenido semántico de una imagen.

Este problema plantea retos técnicos que requieren el uso de arquitecturas avanzadas de aprendizaje profundo. En particular, los modelos *encoder-decoder* han demostrado un desempeño notable al integrar redes neuronales convolucionales (CNN) para la extracción de características visuales y redes recurrentes (RNN) o variantes como LSTM para la generación secuencial de descripciones textuales (Xu et al., 2015; Cornia et al., 2020). La capacidad de tales arquitecturas para mapear representaciones visuales en espacios lingüísticos ha facilitado aplicaciones de alto impacto, como asistentes para personas con discapacidad visual, sistemas de indexación semántica de imágenes y motores de búsqueda visual con capacidades lingüísticas (Hossain et al., 2019; Allamanis et al., 2016).

Sin embargo, esta integración multimodal no está exenta de desafíos. La calidad de las descripciones generadas depende de múltiples factores: desde el diseño de la arquitectura hasta el conjunto de datos utilizado y las métricas de evaluación aplicadas (Hossen et al., 2024). En ese

contexto, esta tesis doctoral presenta un cuerpo de trabajo dividido en tres estudios complementarios que exploran los fundamentos, limitaciones y oportunidades en la descripción automática de imágenes mediante técnicas de *deep learning*.

1.2. Planteamiento del problema

1.2.1. Definición del problema

La generación automática de descripciones de imágenes se ha beneficiado de arquitecturas *encoder-decoder* basadas en aprendizaje profundo. Sin embargo, persisten limitaciones técnicas y metodológicas críticas. Entre ellas destacan: (i) la falta de consenso sobre qué configuraciones arquitectónicas ofrecen un equilibrio óptimo entre precisión y eficiencia computacional, (ii) la dependencia de grandes volúmenes de datos etiquetados, y (iii) la ausencia de métricas que evalúen adecuadamente la calidad sintáctica y semántica de los subtítulos generados (Zhou et al., 2020).

Además, el rendimiento de los modelos *encoder-decoder* se ve afectado significativamente por la elección de optimizadores y el ajuste de hiperparámetros. A pesar del uso extendido de algoritmos como Adam o RMSprop, no existe una guía sobre su desempeño relativo en tareas de clasificación visual ni en su impacto sobre la calidad lingüística de las descripciones generadas.

1.2.2. Delimitación de la investigación

Este trabajo se restringe al estudio de modelos de aprendizaje profundo supervisado para tareas de descripción automática de imágenes. Se abordan tres líneas complementarias: (i) el análisis experimental de optimizadores en redes CNN para clasificación binaria, (ii) la revisión sistemática de la literatura reciente sobre arquitecturas *encoder-decoder*, y (iii) el desarrollo y evaluación de un sistema de reconocimiento facial basado en *deep learning*.

La tesis se centra en modelos con codificadores CNN y decodificadores LSTM, excluyendo técnicas no supervisadas, modelos exclusivamente transformadores y arquitecturas generativas adversariales. Asimismo, el análisis se enfoca en conjuntos de datos públicos y etiquetados como MS COCO, Flickr30K y LFW.

1.3. Preguntas de investigación e hipótesis

Pregunta de investigación:

¿Cuáles son las arquitecturas, algoritmos de optimización y condiciones de entrenamiento que permiten mejorar la calidad sintáctica, semántica y eficiencia computacional de los modelos de descripción automática de imágenes basados en aprendizaje profundo?

Hipótesis central:

La integración de modelos encoder-decoder que emplean CNN como codificadores visuales y LSTM como decodificadores lingüísticos, junto con la selección adecuada de optimizadores como Adam y mecanismos de atención adaptativa, permite generar descripciones automáticas con mayor precisión semántica y estabilidad de entrenamiento en comparación con enfoques tradicionales.

1.4. Objetivo general

Analizar, desarrollar y evaluar estrategias basadas en aprendizaje profundo para la generación automática de descripciones de imágenes, mediante el estudio de arquitecturas *encoder-decoder*, técnicas de optimización y su aplicación a problemas reales como la clasificación visual y el reconocimiento facial.

1.5. Objetivos específicos

1. Analizar el impacto de distintos algoritmos de optimización en redes convolucionales aplicadas a tareas de clasificación, identificando ventajas y limitaciones en escenarios prácticos
2. Realizar una revisión sistemática de la literatura reciente sobre modelos encoder-decoder aplicados a la descripción automática de imágenes, identificando tendencias, métricas y arquitecturas predominantes

3. Diseñar, implementar y evaluar un sistema basado en aprendizaje profundo para el reconocimiento y la descripción de imágenes, validando su desempeño en condiciones adversas y proponiendo lineamientos metodológicos para futuros desarrollos.

1.6. Justificación

La generación automática de descripciones de imágenes constituye una importante línea de investigación en la construcción de sistemas inteligentes capaces de comprender el contenido visual y expresarlo en lenguaje humano. Esta tarea es interdisciplinaria y plantea retos tanto en visión por computadora como en procesamiento de lenguaje natural (Anderson et al., 2018; Hossen et al., 2024).

Desde el punto de vista científico, esta tesis ofrece una contribución significativa al combinar análisis, revisión crítica y desarrollo metodológico en torno al diseño y entrenamiento de modelos *encoder-decoder*. En términos aplicados, los hallazgos permiten construir modelos robustos, escalables y precisos, con impacto directo en áreas como accesibilidad, vigilancia inteligente, búsqueda semántica de imágenes y sistemas autónomos.

El aporte de este trabajo busca fortalecer la base metodológica para investigadores y desarrolladores que trabajan en el diseño de modelos multimodales en entornos reales, contribuyendo al avance del estado del arte en inteligencia artificial explicable y eficiente.

1.7. Metodología utilizada

La generación automática de descripciones de imágenes constituye una línea de investigación esencial en la construcción de sistemas inteligentes capaces de comprender el contenido visual y expresarlo en lenguaje humano. Esta tarea es intrínsecamente interdisciplinaria y plantea retos tanto en visión por computadora como en procesamiento de lenguaje natural (Anderson et al., 2018; Hossen et al., 2024).

Desde el punto de vista científico, esta tesis ofrece una contribución significativa al combinar análisis empírico, revisión crítica y desarrollo metodológico en torno al diseño y entrenamiento de modelos *encoder-decoder*. En términos aplicados, los hallazgos permiten construir modelos

más robustos, escalables y precisos, con impacto directo en áreas como accesibilidad, vigilancia inteligente, búsqueda semántica de imágenes y sistemas autónomos.

El carácter integrado y riguroso de este trabajo busca fortalecer la base metodológica para investigadores y desarrolladores que trabajan en el diseño de modelos multimodales en entornos reales, contribuyendo al avance del estado del arte en inteligencia artificial explicable y eficiente.

Bibliografía

Allamanis, M., Peng, H., and Sutton, C. (2016). A convolutional attention network for extreme summarization of source code. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 12222–12230. https://proceedings.mlr.press/v48/allamanis16.html.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086. https://openaccess.thecvf.com/content_cvpr_2018/CameraReady/1163.pdf.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442. https://www.ijcai.org/proceedings/2017/0704.pdf.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Cornia_Meshed-Memory_Transformer_for_Image_Captioning_CVPR_2020_paper.pdf.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36. https://doi.org/10.1145/3295748.

Hossen, M. B., Ye, Z., Abdussalam, A., and Hassan, S. U. (2024). Attribute-driven filtering: A new

attributes predicting approach for fine-grained image captioning. *Engineering Applications of Artificial Intelligence*, 137:109134. 10.1016/j.engappai.2024.109134.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vinyals_Show_and_Tell_2015_CVPR_paper.pdf.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning (ICML)*, pages 2048–2057. https://proceedings.mlr.press/v37/xuc15.pdf.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049. https://doi.org/10.1609/aaai.v34i07.7005.

Portada Interior

Estudio comparativo de optimizadores en el entrenamiento de una red neuronal convolucional en un modelo de reconocimiento binario

Marco López-Sánchez, José Hernández-Torruco, Betania Hernández-Ocaña, Oscar Chávez-Bosquez

En el aprendizaje automático y profundo, la selección del optimizador es un paso crucial, ya que el rendimiento del modelo depende en gran medida del optimizador elegido. Las arquitecturas de redes neuronales profundas suelen tener múltiples capas de representaciones utilizadas para aprender automáticamente a partir de los datos de entrenamiento. La calibración de hiperparámetros desempeña un papel esencial en el proceso de aprendizaje, dado que los valores de los parámetros pueden variar según el conjunto de datos. El objetivo principal de este trabajo fue determinar qué optimizador puede alcanzar la mejor precisión en la fase de entrenamiento de una red neuronal convolucional utilizada para realizar una clasificación binaria, empleando el conjunto de datos público *Dogs vs. Cats*, que consta de un total de 25,000 imágenes. Los optimizadores comparados fueron SGD, Adam y RMSprop. La arquitectura del modelo desarrollado consta de tres capas convolucionales con sus correspondientes capas de *maxpooling*. Como resultado, los experimentos muestran que el optimizador Adam ofrece un mejor rendimiento. Estos resultados sugieren que debe considerarse el uso de Adam para entrenar redes neuronales convolucionales binarias.

Palabras clave: Clasificación de imágenes, optimizador, aprendizaje profundo, redes neuronales convolucionales.

Universidad Juárez Autónoma de Tabasco, División Académica de Ciencias y Tecnologías de la Información

Capítulo 2

Estudio comparativo de optimizadores en el entrenamiento de una red neuronal convolucional en un modelo de reconocimiento binario

La implementación de la visión por computadora ha atraído el interés de los investigadores en el área de la informática, ya que el aprendizaje profundo se ha convertido en un área prometedora dentro del campo de la visión por computadora (Hinton et al., 2006). El aprendizaje profundo utiliza redes neuronales artificiales internamente como el algoritmo principal para realizar predicciones, clasificaciones, entre otras tareas. La clasificación de imágenes cubre una amplia gama de aplicaciones, incluyendo biometría (Jaseena and Kovoor, 2018), vigilancia por video (Ojha and Sakhare, 2015) e investigación médica (Lakhani and Sundaram, 2017).

Entre los algoritmos más prometedores, la técnica de redes neuronales convolucionales (CNN) (LeCun et al., 2015) es el algoritmo de aprendizaje profundo más utilizado. En este trabajo, describimos un problema de clasificación binaria y analizamos los optimizadores Adam, RMSprop y SGD, los cuales se emplean para entrenar redes neuronales de aprendizaje profundo. Adam es el optimizador más reciente y representa una mejora respecto a otros, pero no siempre genera el mejor modelo. A continuación, presentamos conceptos clave y técnicas de optimización, seguidos

de una breve descripción de las redes neuronales convolucionales. El resto del trabajo se organiza de la siguiente manera: la Sección 2 presenta materiales y métodos, la Sección 3 muestra los resultados de los experimentos y, finalmente, la Sección 4 concluye el artículo.

2.0.1. Redes Neuronales Artificiales

Las redes neuronales artificiales (ANN) buscan realizar tareas computacionales utilizando un gran número de unidades de procesamiento interconectadas llamadas neuronas o nodos. Las conexiones entre neuronas están asociadas con parámetros llamados pesos, los cuales pueden modificarse mediante un proceso de entrenamiento para asociar la salida deseada con una entrada específica (Camastra and Vinciarelli, 2015). Se pueden describir como un grafo dirigido en el que cada nodo realiza una función de transferencia (Yao, 1999).

Los hiperparámetros son un conjunto particular de parámetros utilizados en el proceso de aprendizaje durante el entrenamiento. Son fundamentales, ya que afectan directamente la fase de entrenamiento de la red neuronal y, por ende, el desempeño del modelo. Existen varias técnicas de optimización de hiperparámetros, pero es responsabilidad del científico de datos inicializar estos valores manualmente. Los hiperparámetros pueden ser enteros, continuos o categóricos dentro de un rango de valores (Victoria and Maragatham, 2020).

Los hiperparámetros básicos de una ANN son:

- Número de capas ocultas: Las redes neuronales con una sola capa de parámetros ajustables son muy limitadas en lo que pueden hacer, como ocurre con los perceptrones. Por lo tanto, es natural ampliar la capacidad de una red neuronal agregando capas adicionales de neuronas. Desde el exterior, solo son visibles la primera y última capa de una red multicapa: la capa de entrada y la de salida. Las demás capas son ”ocultas” porque no son visibles desde el exterior (Berzal, 2019).

- Número de unidades ocultas: Cada capa oculta puede tener un número diferente de neuronas.

- Tasa de aprendizaje: Es un número pequeño, generalmente entre 0.00001 y 0.05, que determina el tamaño del ajuste aplicado a los pesos actuales durante el entrenamiento del

modelo (Campesato, 2020). En casi todos los algoritmos de descenso de gradiente, la elección de la tasa de aprendizaje es crucial para la eficiencia. (Bengio, 2012) afirma que es uno de los hiperparámetros más importantes y siempre debe ajustarse.

- Número de épocas: Una época significa que cada muestra en el conjunto de datos de entrenamiento ha tenido la oportunidad de actualizar los parámetros internos del modelo. Una época puede estar compuesta por uno o más lotes de datos. Por ejemplo, una época con un solo lote corresponde al algoritmo de aprendizaje por descenso de gradiente por lotes (Brownlee, 2016).

- Tamaño del lote: Es un hiperparámetro que define el número de muestras que se procesan antes de actualizar los parámetros internos del modelo (Brownlee, 2016).

- Función de pérdida: Es una herramienta para medir el desempeño de un algoritmo de decisión (o clasificador). La función de pérdida mide el costo de cada acción del clasificador y convierte un error de probabilidad en una decisión (Camastra and Vinciarelli, 2015).

- Función de activación: Es crucial, ya que las imágenes y sus características son altamente no lineales. La función de activación introduce no linealidad mientras asigna los valores de entrada a su rango correspondiente. Entre las más comunes están Sigmoide, Tanh y ReLU.

El papel de un optimizador es actualizar los pesos de la red neuronal para minimizar la función de pérdida, es decir, la diferencia entre el valor real y el valor predicho. Antes de iniciar la fase de entrenamiento, se debe seleccionar el optimizador adecuado junto con sus hiperparámetros específicos. Los optimizadores comparados en este estudio son:

SGD El algoritmo de descenso de gradiente estocástico (Robbins and Monro, 1951) es uno de los primeros métodos utilizados para entrenar redes neuronales. Su principal ventaja es la simplicidad de cada iteración en la generación de la dirección de búsqueda y la actualización de variables (Nesterov, 2012).

RMSprop Fue introducido para abordar el problema de la tasa de aprendizaje decreciente de forma monótona, presente en el optimizador AdaGrad (Hinton et al., 2012), mediante el uso de una tasa de decaimiento exponencial en el primer paso.

Adam El algoritmo de optimización Adam (Kingma and Ba, 2014) combina los beneficios de AdaGrad y RMSprop.

Las métricas de rendimiento utilizadas en este trabajo son:

- Precisión: Es la métrica más utilizada para resumir el desempeño de un modelo de aprendizaje supervisado, ya que indica la proporción de ejemplos clasificados correctamente (Berzal, 2019).
- Matriz de confusión: Es una tabla de contingencia que contiene los valores de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos (Campesato, 2020).
- Curva ROC: Representa gráficamente la tasa de verdaderos positivos (recall) frente a la tasa de falsos positivos (Campesato, 2020).

2.0.2. Redes Neuronales Convolucionales

Una red neuronal convolucional (CNN) es un algoritmo de aprendizaje profundo especializado en la clasificación de imágenes. Su núcleo es el procesamiento de datos mediante la operación de convolución (El-Amir and Hamdy, 2019). LeCun et al. (1990) propusieron el marco moderno de las CNN y lo mejoraron posteriormente (LeCun et al., 2015).

Cuando se aplican CNN a imágenes, las capas convolucionales sucesivas aprenden características progresivamente más abstractas. La Figura 2.1 ilustra este proceso.

2.0.3. Conjunto de datos

El conjunto de datos Dogs vs. Cats fue recopilado originalmente por Microsoft y consta de más de 3 millones de imágenes, de las cuales 25,000 han sido publicadas públicamente[1]. Vale la pena mencionar que existe un conjunto de datos más pequeño (3,000 imágenes) derivado de este, que es más comúnmente utilizado en la literatura.

Los datos son imágenes, que están almacenadas en 2 directorios, uno para las imágenes de gatos y otro para las imágenes de perros. El conjunto de datos está equilibrado, es decir,

[1] https://www.microsoft.com/en-us/download/confirmation.aspx?id=54765

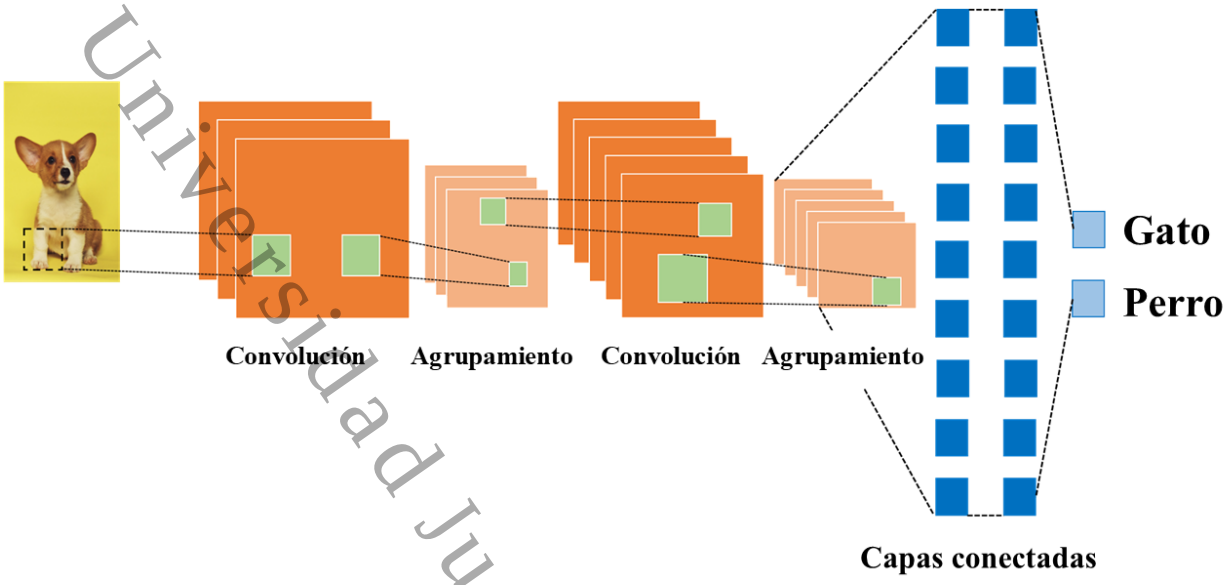

Figura 2.1. Ejemplo de una red neuronal convolucional.

tiene la misma cantidad de imágenes de gatos y perros. El objetivo es construir un algoritmo de aprendizaje automático capaz de detectar correctamente el animal (perro o gato) en las imágenes.

2.0.4. Configuración experimental

El modelo de CNN propuesto en este trabajo se utiliza para predecir la clase correcta de cada elemento en el conjunto de datos. Los datos se dividieron en 70 % para entrenamiento, 15 % para validación y 15 % para prueba. El número total de capas convolucionales es 3*stacks. Cada stack tiene una capa convolucional, una capa de normalización por lotes (batch normalization) y una capa ReLU. El tamaño de la imagen es $150*150*3$, donde $150*150$ es el tamaño en píxeles, y el número 3 representa la profundidad de la imagen.

La configuración de hiperparámetros fue:

- Tasa de aprendizaje: 0.00001
- Número de épocas: 30
- Pasos por época: 70
- Función de pérdida: Entropía cruzada binaria
- Funciones de activación: ReLU y Sigmoide

El ciclo de vida de un modelo proporciona la estructura para modelar un conjunto de datos. El ciclo de vida utilizado en este trabajo incluye (Rosebrock, 2017):

1. Recolectar el conjunto de datos: descargar un archivo zip que contiene el conjunto de datos, descomprimirlo y descartar los archivos corruptos.

2. Dividir el conjunto de datos: dividir el conjunto de datos en tres partes:

   - Conjunto de entrenamiento: el conjunto de muestras utilizado para entrenar el modelo.
   - Conjunto de validación: el conjunto de muestras que se usa para ajustar los parámetros del clasificador.
   - Conjunto de prueba: el conjunto de muestras que se usará únicamente para evaluar el rendimiento del clasificador.

Se asignó el 70 % de las imágenes al conjunto de entrenamiento, el 15 % al conjunto de validación y el 15 % al conjunto de prueba.

Entrenar la red: En este paso, se define la arquitectura de la red configurando cada una de sus capas. También se configuran los hiperparámetros, se compila el modelo y se alimenta con el conjunto de entrenamiento.

Evaluación: Para verificar el resultado obtenido en el entrenamiento, se utiliza la matriz de confusión y la curva ROC.

Realizar predicciones: Se proporciona una interfaz de usuario simple para que el usuario pueda cargar una imagen y probar la predicción de manera más intuitiva.

2.1. Resultados

Se realizaron experimentos con el conjunto de datos para determinar el comportamiento de cada optimizador. Para evaluar el rendimiento de cada optimizador en todos los experimentos, se eligieron las configuraciones predeterminadas de cada hiperparámetro. Se incluyó una detención temprana en la fase de entrenamiento para detener el ciclo de entrenamiento después de 5 iteraciones sin mejora. La Tabla 2.1 muestra que la mayor precisión se logró con el optimizador

Tabla 2.1. Resultados del modelo.

Optimizador	Pérdida	Precisión
SGD	0.6926	0.5053
RMSprop	0.4520	0.7868
Adam	0.4227	0.8017

Adam, mientras que RMSprop quedó en segundo lugar con una pequeña diferencia. SGD tuvo un desempeño inferior con una precisión débil.

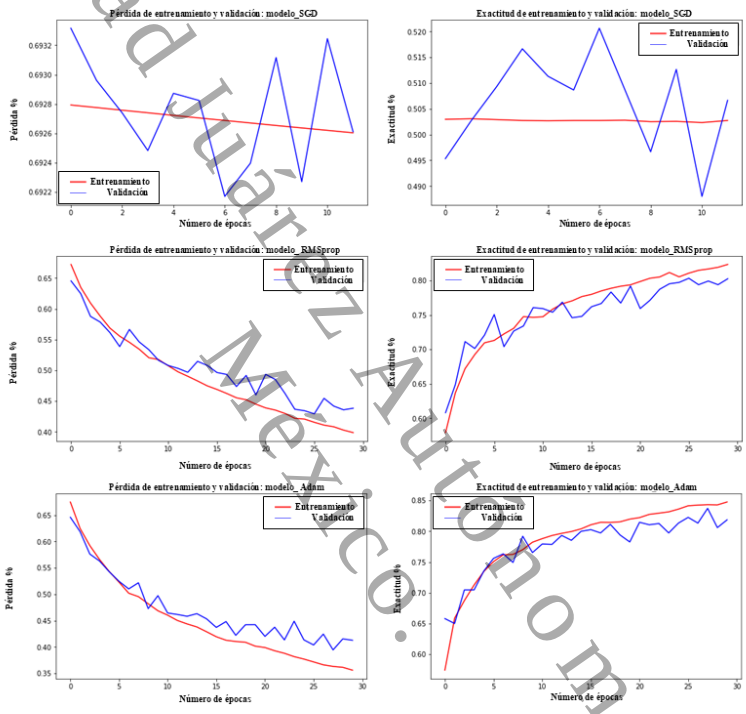

Figura 2.2. Rendimiento de los optimizadores.

La Figura 2.2 muestra el rendimiento de cada modelo en la fase de entrenamiento en los subconjuntos de entrenamiento y validación. Se destaca que SGD detiene el entrenamiento en la época 12, según la configuración de detención temprana. Por otro lado, RMSprop y Adam continúan el entrenamiento hasta las 30 épocas configuradas. La curva de validación de SGD muestra un comportamiento errático, mientras que las curvas de RMSprop y Adam muestran la pendiente del proceso de aprendizaje.

La matriz de confusión de cada optimizador se presenta en la Figura 2.3. Se puede notar que SGD tiene el peor rendimiento, clasificando incorrectamente la mayoría de las imágenes. Por otro

lado, RMSprop y Adam lograron una clasificación bastante precisa de gatos y perros.

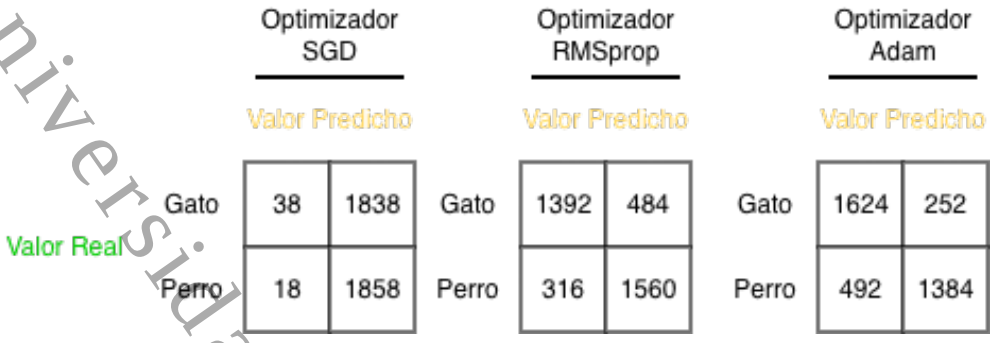

Figura 2.3. Matriz de confusión de cada optimizador.

En la Figura 2.4 se puede ver el área bajo la curva (AUC) de la curva ROC de cada optimizador, calculando la tasa de falsos positivos y la tasa de verdaderos positivos de cada modelo. El AUC mide qué tan bien el modelo es capaz de distinguir entre gatos y perros. Se observa que la curva del optimizador SGD es plana, lo que indica un bajo rendimiento en la clasificación. La curva ROC de RMSprop y Adam es similar, al igual que su AUC. Ambos optimizadores lograron un buen desempeño en la tarea de clasificación.

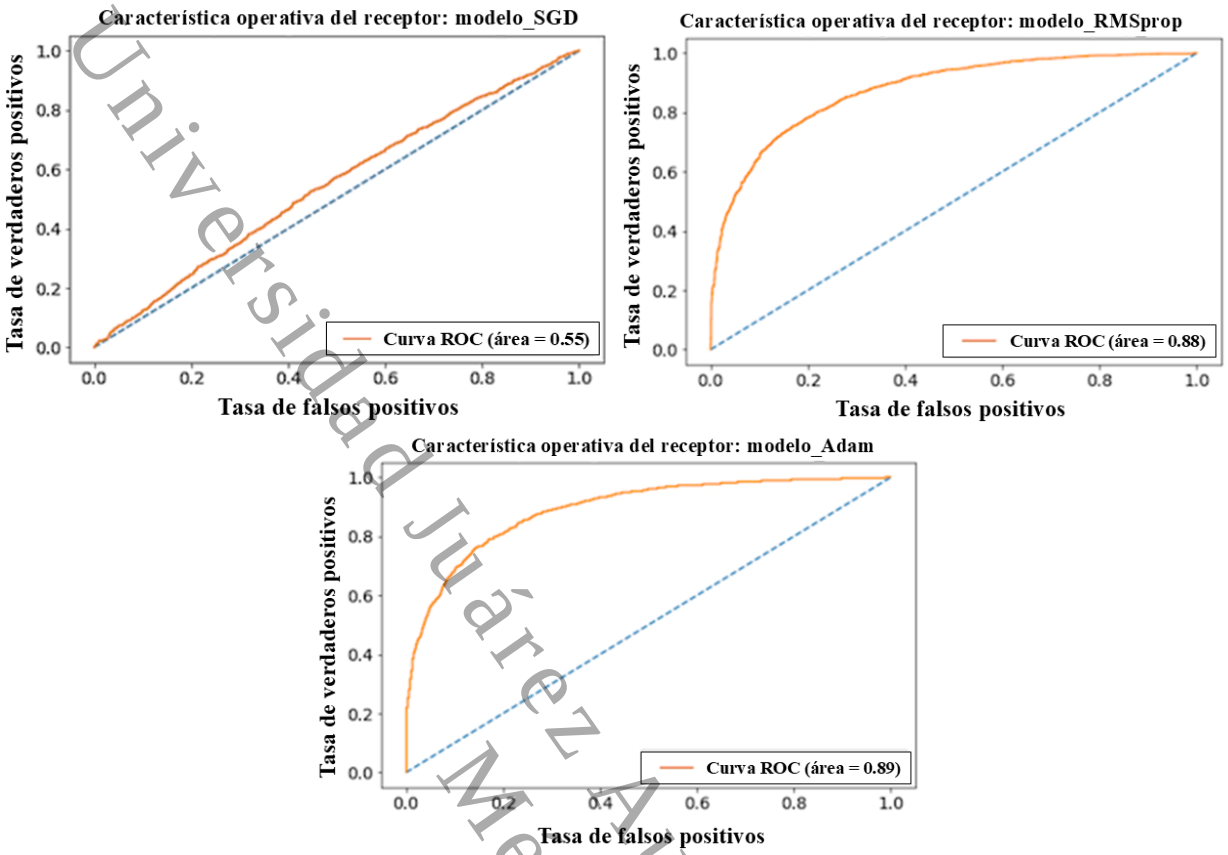

Figura 2.4. Curva ROC de cada optimizador.

2.2. Conclusiones

Seleccionar un optimizador para entrenar una red neuronal es una tarea desafiante. En este trabajo, se probaron tres optimizadores: SGD, RMSprop y Adam. En nuestros experimentos, Adam obtuvo el mejor rendimiento. En futuros trabajos, consideraremos la optimización de parámetros y el uso de optimizadores más diversos.

Bibliografía

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer.

Berzal, F. (2019). *Redes neuronales & Deep Learning: Volumen II*. Edición Independiente, 1era. edition.

Brownlee, J. (2016). *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. Jason Brownlee, 1st edition.

Camastra, F. and Vinciarelli, A. (2015). *Machine learning for audio, image and video analysis: theory and applications*. Springer. url=https://doi.org/10.1007/978-1-4471-6735-8.

Campesato, O. (2020). *Artificial Intelligence, Machine Learning, and Deep Learning*. Stylus Publishing, LLC.

El-Amir, H. and Hamdy, M. (2019). *Deep Learning Pipeline: Building a Deep Learning Model with TensorFlow*. Apress. https://link.springer.com/book/10.1007/978-1-4842-5349-6.

Hinton, G., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527.

Hinton, G., Srivastava, N., and Swersky, K. (2012). Neural networks for machine learning lecture: Lecture 6a overview of mini-batch gradient descent. Coursera. https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Jaseena, K. and Kovoor, B. (2018). A survey on deep learning techniques for big data in biometrics. *International Journal of Advanced Research in Computer Science*, 9(1):12–17. https://doi.org/10.26483/ijarcs.v9i1.5136.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://arxiv.org/pdf/1412.6980.

Lakhani, P. and Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582. https://pubmed.ncbi.nlm.nih.gov/28436741/.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. https://doi.org/10.1038/nature14539.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in*

neural information processing systems, pages 396–404. https://dl.acm.org/doi/10.5555/109230.109279.

Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362. https://doi.org/10.1137/100802001.

Ojha, S. and Sakhare, S. (2015). Image processing techniques for object tracking in video surveillance-a survey. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–6. IEEE. https://materias.df.uba.ar/l5a2021c1/files/2021/05/ojha2015.pdf.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407. https://www.columbia.edu/~ww2040/8100F16/RM51.pdf.

Rosebrock, A. (2017). *Deep Learning for Computer Vision with Python: Starter Bundle*. PyImageSearch.

Victoria, A. H. and Maragatham, G. (2020). Automatic tuning of hyperparameters using bayesian optimization. *Evolving Systems*. https://doi.org/10.1007/s12530-020-09345-2.

Yao, X. (1999). Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447. https://doi.org/10.1109/5.784219.

Portada Interior

Técnicas de Aprendizaje Profundo Supervisado para la Descripción de Imágenes: Una Revisión Sistemática

Marco López-Sánchez, Betania Hernández-Ocaña, Oscar Chávez-Bosquez, José Hernández-Torruco

La descripción automática de imágenes, también conocida como *image captioning*, tiene como objetivo describir los elementos presentes en una imagen y las relaciones entre ellos. Esta tarea involucra dos campos de investigación: la visión por computadora y el procesamiento del lenguaje natural; por ello, ha recibido una gran atención dentro de la ciencia de la computación.
En este artículo de revisión, seguimos la metodología de revisión de Kitchenham para presentar los enfoques más relevantes sobre las metodologías de descripción de imágenes basadas en aprendizaje profundo. Nos centramos en los trabajos que utilizan redes neuronales convolucionales (CNN) para extraer las características de las imágenes y redes neuronales recurrentes (RNN) para la generación automática de oraciones.
Como resultado, se seleccionaron 53 artículos de investigación que emplean el enfoque *encoder-decoder*, enfocándose únicamente en aprendizaje supervisado. Las principales contribuciones de esta revisión sistemática son:
(i) describir los trabajos más relevantes sobre descripción de imágenes que implementan un enfoque *encoder-decoder* entre los años 2014 y 2022, y
(ii) determinar las principales arquitecturas, conjuntos de datos y métricas que se han aplicado en la tarea de descripción de imágenes.

Palabras clave: Descripción de imágenes (*image captioning*); visión por computadora; procesamiento del lenguaje natural; red neuronal convolucional; red neuronal recurrente.

Universidad Juárez Autónoma de Tabasco, División Académica de Ciencias y Tecnologías de la Información

Capítulo 3

Técnicas de Aprendizaje Profundo Supervisado para la Descripción de Imágenes: Una Revisión Sistemática

3.1. Introducción

La capacidad de generar descripciones automáticas de imágenes conecta dos campos: visión por computadora y procesamiento del lenguaje natural, ambos destacados en la informática. La visión por computadora es necesaria para extraer las características de las imágenes, mientras que las técnicas de procesamiento del lenguaje natural ayudan a convertir esas características en una descripción adecuada para los humanos. Debido a la relación entre ambas disciplinas, la descripción de imágenes está fuertemente vinculada a los avances en ambos campos, encontrando en el aprendizaje profundo un punto de unión.

El uso de modelos de aprendizaje profundo para generar descripciones automáticas de imágenes incluye la implementación de redes neuronales convolucionales (CNN) para realizar la parte de visión artificial y la extracción de características. Además, la mayoría de los modelos basados en aprendizaje profundo utilizan redes neuronales recurrentes (RNN) para realizar tareas de procesamiento del lenguaje natural. Por esta razón, la generación automática de subtítulos de imágenes está actualmente vinculada al conjunto de técnicas y arquitecturas de aprendizaje

profundo; por lo tanto, este trabajo se centrará en estas. Además, otras técnicas de aprendizaje profundo se utilizan para resolver este problema, como las redes generativas adversarias (GANs) (Wang et al., 2018; Dai et al., 2017; Shetty et al., 2017; Mohamad Nezami et al., 2019; Jiang et al., 2021). Las GANs son un método emergente de aprendizaje semisupervisado y no supervisado; por lo tanto, no se incluyen en esta revisión.

Este artículo está organizado de la siguiente manera: La Sección 3.2 describe los antecedentes y trabajos relacionados. La Sección 3.3 presenta la metodología. La Sección 3.4 se centra en los resultados y discusiones respecto a las preguntas de investigación. Finalmente, la Sección 3.5 presenta la conclusión.

3.2. Antecedentes

En esta sección, primero esbozamos tres algoritmos de aprendizaje profundo identificados en la literatura, introducimos los conceptos básicos de la descripción de imágenes y describimos la arquitectura general codificador-decodificador. Esta arquitectura se basa en métodos de aprendizaje profundo para la descripción de imágenes.

3.2.1. Descripción de Imágenes

La Asociación Americana de Antropología define *descripción de imágenes* como “una explicación detallada de una imagen que proporciona acceso textual al contenido visual; se usa con mayor frecuencia para gráficos digitales en línea y en archivos digitales” (The American Anthropological Association, 2019). La descripción automática de imágenes, también conocida como generación de subtítulos de imágenes o *image captioning*, es el proceso de generar una descripción concisa y comprensible para los humanos sobre el contenido de una imagen. Esta oración en lenguaje natural debe describir los objetos, entidades y relaciones de manera similar a como lo haría una persona (Amirian et al., 2020).

Por lo tanto, la descripción precisa de imágenes es una tarea desafiante que requiere tecnología de última generación en visión por computadora y procesamiento del lenguaje natural. La Figura 3.1 muestra algunas imágenes y sus correspondientes subtítulos generados automáticamente.

(a) (b) (c)

(d) (e) (f)

Figura 3.1. Ejemplos de subtítulos generados por modelos de descripción automática de imágenes. (**a**) Una persona sosteniendo una caja de pizza (Zhang et al., 2017). (**b**) Una señal de alto en una carretera con una montaña al fondo (Xu et al., 2015). (**c**) Una mesa de madera y sillas organizadas en una habitación (Kiros et al., 2014a). (**d**) Cinco personas de pie y cuatro en cuclillas sobre una roca marrón en primer plano Mao et al. (2014). (**e**) Un hombre con una camisa negra tocando una guitarra (Wang and Chan, 2018). (**f**) Un grupo de jugadores de béisbol jugando un partido (Chen and Lawrence Zitnick, 2015).

Como un tema emergente en el aprendizaje profundo, hay aplicaciones prometedoras para la descripción automática de imágenes, tales como:

- **Generación de texto alternativo para personas con discapacidad visual.** Las personas con ceguera o baja visión pueden comprender imágenes en páginas web o escenas del mundo real mediante la conversión automática de una imagen en texto y su descripción utilizando un sistema de conversión de texto a voz. Esta técnica puede permitir a las personas con discapacidad visual obtener la mayor cantidad de información posible sobre el contenido de una fotografía.

- **Recuperación de imágenes basada en contenido (CBIR).** Consiste en recuperar un subconjunto específico de imágenes (o una imagen individual) a partir de palabras clave relevantes que reflejen el contenido visual de la imagen. CBIR y los enfoques de extracción de características se aplican en diversas aplicaciones, como análisis de imágenes médicas,

teledetección, detección de delitos, análisis de video, vigilancia militar y la industria textil.

Actualmente, el enfoque estándar para la descripción automática de imágenes consiste en una implementación codificador-decodificador utilizando técnicas de aprendizaje profundo para extraer e interpretar las características de una imagen y generar una oración que describa la escena. Las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN) suelen ser los protagonistas del enfoque codificador-decodificador.

En términos generales, el elemento fundamental para el funcionamiento de las redes neuronales es la información, medida mediante la entropía. Las redes neuronales intentan preservar la mayor cantidad posible de información de entrada a medida que esta pasa entre las diferentes capas y comprimirla para optimizar su rendimiento. Diferentes arquitecturas de redes neuronales abordan este problema con su propia estrategia, dependiendo de la tarea que desean realizar. Por ejemplo, las Redes Neuronales Convolucionales se especializan en el reconocimiento y clasificación de imágenes, mientras que las Redes Neuronales Recurrentes resuelven tareas de procesamiento del lenguaje natural, como la traducción de texto o el análisis de opiniones y discursos.

3.2.2. Redes Neuronales Convolucionales

Las Redes Neuronales Convolucionales (CNN) son un algoritmo de aprendizaje profundo especializado en la clasificación de imágenes. El elemento central de una CNN es el procesamiento de datos mediante la operación de convolución. En 1990, (LeCun et al., 1990) publicaron el artículo seminal que sentó las bases del marco moderno de una CNN y posteriormente lo mejoraron en (LeCun et al., 2015). Una CNN se especializa en emular la funcionalidad y el comportamiento de nuestra corteza visual (Sarkar et al., 2018).

La Figura 3.2 muestra una CNN típica que incluye tres tipos de capas:

- Capa de convolución: La extracción de características se realiza mediante filtros llamados kernels, cada uno generalmente seguido por una capa ReLU.
- Capa de agrupamiento (pooling): Se aplica un barrido para obtener información estadística, reduciendo así el vector que representa la imagen procesada.

- Capa de aplanamiento (flattening): Finalmente, se aplica una capa de aplanamiento para transformar la matriz en un vector unidimensional; el vector resultante será el que alimentará a la red neuronal para realizar tareas de detección o clasificación.

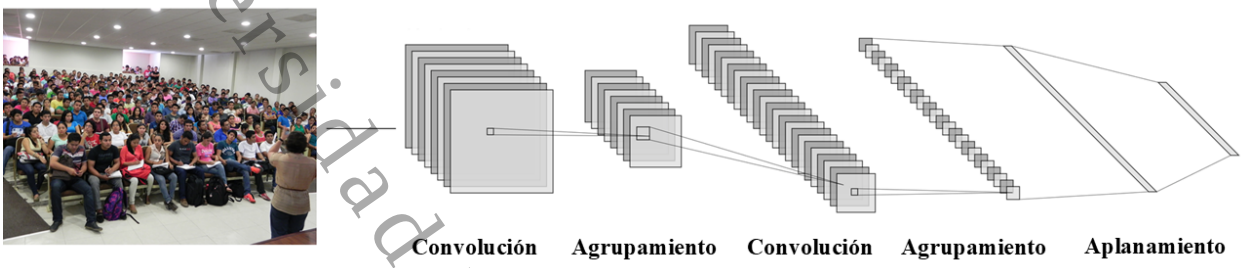

Figura 3.2. Arquitectura típica de una CNN.

3.2.3. Redes Neuronales Recurrentes

Las Redes Neuronales Recurrentes (RNN) (Pascanu et al., 2014) han sido ampliamente utilizadas en el aprendizaje profundo. Las RNN surgen como una solución al problema del aislamiento de la información a lo largo del tiempo. Este es un problema para ciertos datos que dependen de sus predecesores, como el texto, ya que analizar cada palabra de forma independiente provoca una pérdida de información importante. Para superar esta limitación, las RNN permiten conexiones hacia atrás en la red, alimentándose de la información previamente procesada como una especie de memoria. Las RNN se utilizan extensamente en visión por computadora, particularmente en la generación automática de descripciones de imágenes utilizando el modelo codificador-decodificador (Vinyals et al., 2015). En un modelo RNN, a medida que se agregan más capas a la red neuronal, los gradientes de la función de pérdida tienden a cero, lo que dificulta el entrenamiento de la red. Esto se conoce como el efecto del “gradiente que se desvanece”.

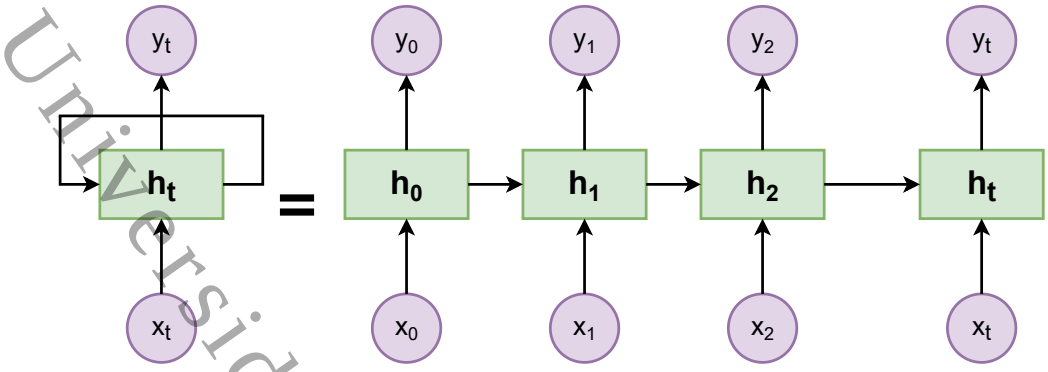

Figura 3.3. Arquitectura típica de una RNN.

Una RNN simula un sistema dinámico en tiempo discreto que tiene una entrada $\mathbf{x}_t$, una salida $\mathbf{y}_t$ y un estado oculto $\mathbf{h}_t$. En la Figura 3.3, el subíndice $\mathbf{t}$ representa el tiempo. La arquitectura de una Red Neuronal Recurrente es una secuencia de redes neuronales que están conectadas una tras otra mediante retropropagación. La Figura 3.3 ilustra una RNN desplegada. En el lado izquierdo, la RNN está desplegada después del signo igual; se visualizan los diferentes pasos temporales, y la información se transfiere de un paso temporal al siguiente (Pascanu et al., 2014).

3.2.4. Memoria a Largo Plazo (LSTM)

LSTM es una variante del modelo RNN propuesta para abordar el problema del gradiente que se desvanece. Esta arquitectura presenta una celda de memoria que permite mantener su estado a lo largo del tiempo, respaldada por unidades conocidas como puertas. La configuración de LSTM más comúnmente utilizada en la literatura se denomina Vanilla LSTM (Greff et al., 2017). La Figura 3.4 muestra la arquitectura de un bloque típico de LSTM vanilla.

Los elementos principales de una LSTM son (Houdt et al., 2020):

- **Entrada del bloque:** actualiza el componente de entrada del bloque, que combina la entrada actual $x^{(t)}$ y la salida de esa unidad LSTM $y^{(t-1)}$ en la iteración anterior.
- **Puerta de entrada:** combina la entrada actual $x^{(t)}$, la salida de esa unidad LSTM $y^{(t-1)}$ y $c^{(t-1)}$ en la iteración anterior.
- **Puerta de olvido:** la unidad LSTM determina qué información debe eliminarse de sus estados anteriores de celda $c^{(t-1)}$.

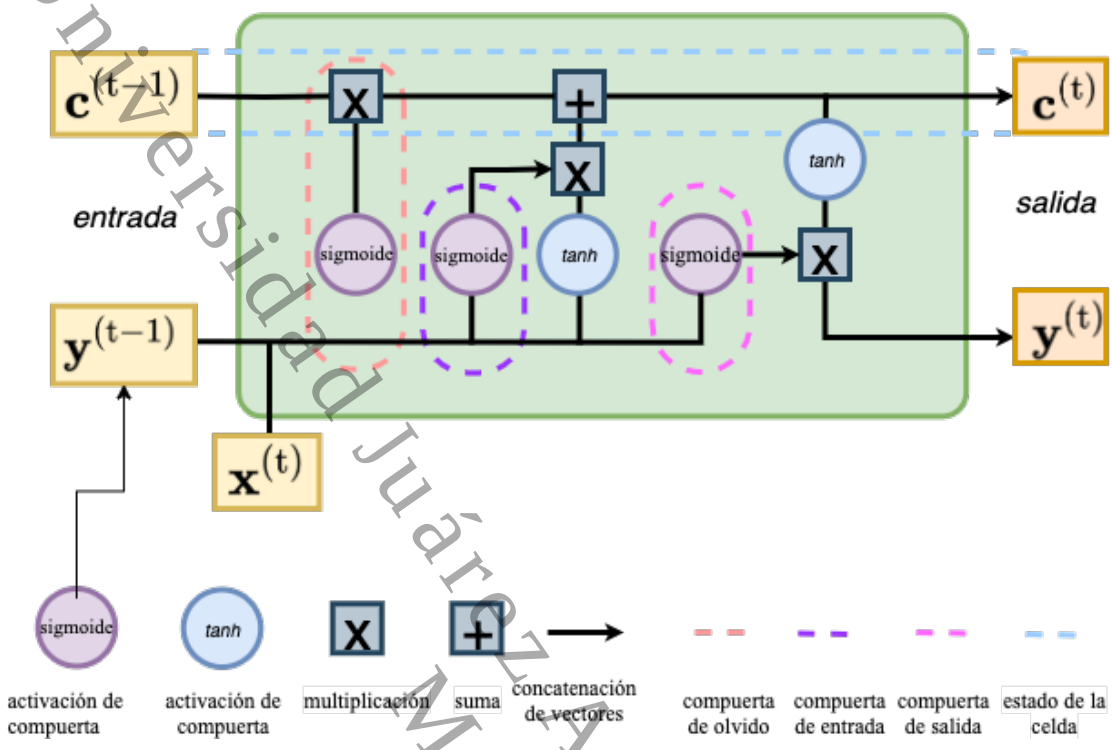

Figura 3.4. Arquitectura típica de un bloque LSTM.

- **Celda:** calcula el valor de la celda, que combina la entrada del bloque $z^{(t)}$, la puerta de entrada $i^{(t)}$, y la puerta de olvido $f^{(t)}$ con el valor anterior de la celda.
- **Puerta de salida:** calcula la puerta de salida, que combina la entrada actual $x^{(t)}$, la salida de esa unidad LSTM $y^{(t-1)}$ y el valor de la celda $c^{(t-1)}$ en la iteración anterior.
- **Bloque de salida:** combina el valor actual de la celda $c^{(t)}$ con el valor actual de la puerta de salida.

3.2.5. Enfoque Codificador-Decodificador

La arquitectura codificador-decodificador nació para la traducción automática. En esta arquitectura, una red codificadora codifica una frase en algún idioma como un vector de longitud fija. Luego, otra red decodificadora lee el vector codificado y genera una secuencia de salida en un nuevo idioma.

Primero, el enfoque codificador-decodificador utiliza un modelo de aprendizaje profundo para

codificar la imagen en un vector de características. A continuación, el modelo decodificador utiliza el vector de entrada para generar una oración en lenguaje natural que describe la imagen (Kiros et al., 2014b). Este es el enfoque más común para abordar la descripción de imágenes, dado los resultados prometedores para esta tarea. Las CNN han sido y continúan siendo la arquitectura de red más utilizada para la codificación de imágenes y la extracción de características. En contraste, las RNN tienen la función de decodificar estas características en oraciones, es decir, en la descripción de la imagen (Fang et al., 2015). Ambos modelos se entrenan conjuntamente en la arquitectura codificador-decodificador para maximizar la probabilidad de la oración dada la imagen (Karpathy and Fei-Fei, 2015). La Figura 3.5 proporciona una visión general de los conceptos básicos y del mecanismo de un generador automático de descripciones de imágenes (Amirian et al., 2020).

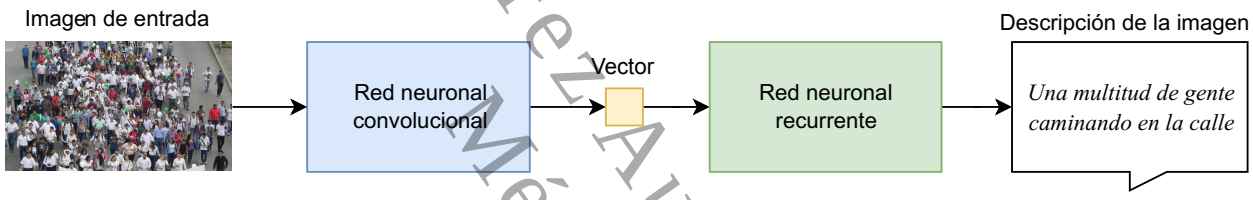

Figura 3.5. Una arquitectura general para la descripción de imágenes utilizando aprendizaje profundo.

Una canalización típica de codificador-decodificador incluye la extracción, el filtrado y la transformación de pares imagen/subtítulo para obtener un modelo preciso para la descripción automática de imágenes.

Los siguientes pasos representan un flujo de trabajo mínimo para entrenar modelos de descripción automática de imágenes:

1. Seleccionar un conjunto de datos. Es necesario utilizar un conjunto de datos que incluya una gran colección de imágenes (en el orden de mil imágenes), cada una con varios subtítulos que proporcionen una descripción precisa de su contenido.

2. Codificador (modelo de extracción de características). Las CNN son la herramienta por defecto para extraer las características de la imagen de entrada. Una CNN realiza una reducción de dimensionalidad, donde los píxeles de la imagen se representan de tal manera que las partes interesantes de la figura se capturan eficazmente en señales de características extraídas. Actualmente, esta tarea puede abordarse de una de las siguientes formas:

- Entrenar la CNN directamente con las imágenes del conjunto de datos de subtitulación de imágenes;
- Utilizar un modelo de clasificación de imágenes preentrenado, como el modelo VGG, ResNet50, Inception V3 o EfficientNetB7.

3. Las señales de características extraídas se representan en un vector de codificación de longitud fija. Este vector contiene una representación rica de la imagen de entrada.

4. Decodificador (modelo de lenguaje). Las RNN son la herramienta por defecto para trabajar con problemas de predicción de secuencias. En esta etapa, la RNN toma el vector de codificación de longitud fija y predice la probabilidad de la siguiente palabra en una secuencia, dada una lista de palabras presentes en dicha secuencia. La salida es entonces la descripción en lenguaje natural de la imagen. Actualmente, las redes LSTM son una arquitectura de RNN comúnmente utilizada, ya que permiten encapsular una secuencia más amplia de palabras u oraciones que las RNN convencionales.

3.3. Metodología

El método de revisión sistemática de la literatura se basa en la metodología propuesta por Kitchenham (2004). La Figura 3.6 muestra el proceso de revisión sistemática utilizado en este trabajo, que consiste en tres fases:

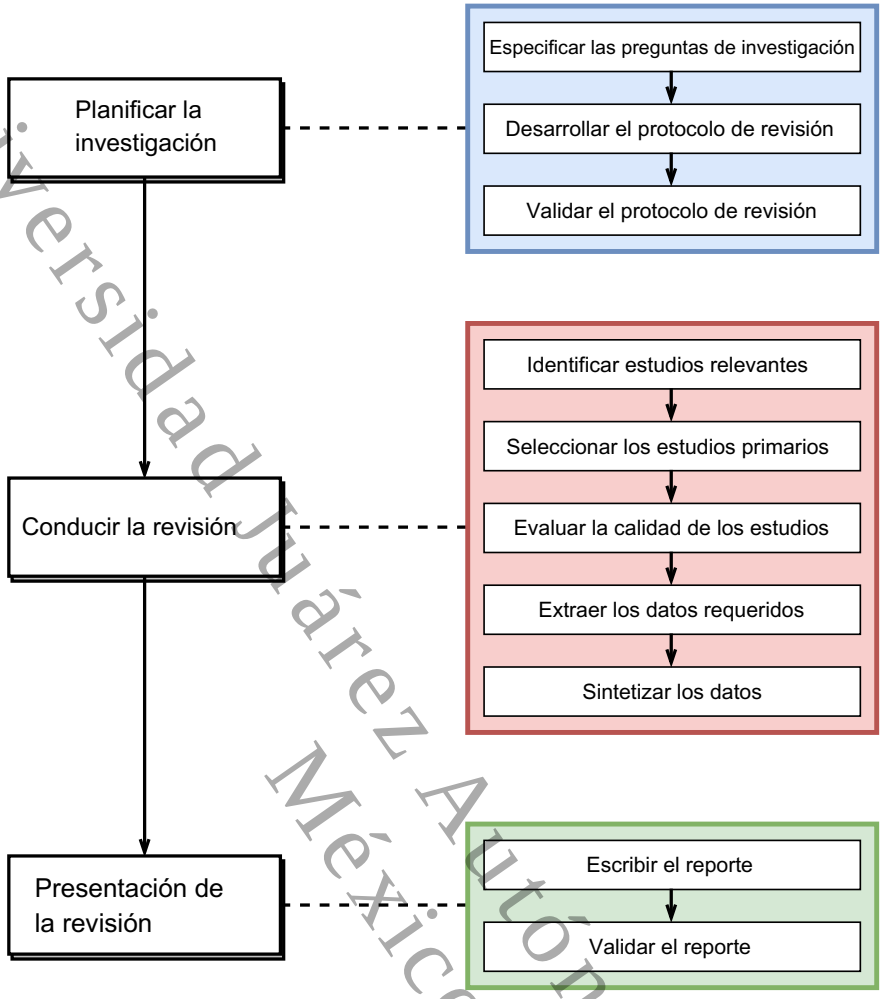

Figura 3.6. Proceso de revisión sistemática.

1. *Planificación de la investigación.* La primera fase consiste en formular las preguntas de investigación adecuadas para identificar el tema de estudio. Este trabajo examina las siguientes preguntas de investigación (RQ):

   - RQ1: ¿Cuál es la arquitectura personalizada implementada en el enfoque codificador-decodificador utilizado en los artículos de investigación?
   - RQ2: ¿Qué conjuntos de datos se utilizan para entrenar y probar los modelos?
   - RQ3: ¿Qué métricas se utilizaron para evaluar los resultados obtenidos?

   Estas tres preguntas forman la base para desarrollar una estrategia de investigación para la extracción de literatura.

Luego de definir las preguntas de investigación, la siguiente actividad en la fase de planificación consiste en la selección de fuentes para definir la estrategia de búsqueda. Para este estudio, se seleccionaron las siguientes bases de datos bibliográficas internacionales en línea:

- IEEE Xplore.
- ACM Digital Library.
- ScienceDirect.
- Springer.

Las búsquedas se limitaron a publicaciones revisadas por pares, escritas en inglés y publicadas entre 2014 y 2022. Las palabras clave utilizadas para las búsquedas fueron: “Encoder-decoder for automatic image description”, “Encoder-decoder for automatic image captioning”, “deep learning for image description”, y “Evaluation of image description generator models”. El último paso en la fase de planificación fue la selección y evaluación de los artículos de investigación. Durante esta fase, se realizó una selección inicial que involucró la revisión de títulos, palabras clave y resúmenes de los posibles estudios primarios.

Esta primera fase nos permite identificar la perspectiva actual del problema de investigación, definir los modelos y enfoques más recientes utilizados para resolverlo, y confirmar que se trata de un tema de interés actual.

2. *Ejecución de la búsqueda*. La segunda fase relacionada con la metodología consiste en realizar la extracción y análisis de documentos de las bases de datos en línea. Se aplicaron los siguientes requisitos para asegurar que los hallazgos se clasificaran adecuadamente:

   - El diseño e implementación de un modelo de aprendizaje profundo para la descripción de imágenes es el tema central que propone este estudio.
   - Los estudios primarios reportan todos los componentes esenciales sobre los que se construye un enfoque codificador-decodificador para la descripción de imágenes.
   - Los estudios primarios reportan todas las métricas utilizadas para evaluar el modelo de descripción de imágenes.

- Los artículos de investigación mencionan los conjuntos de datos empleados.

Para obtener una lista concisa de artículos, se realizó una verificación comparativa para detectar artículos duplicados. Además, fue obligatorio analizar las secciones de introducción y conclusión para determinar qué artículos seleccionar o descartar.

Después de analizar y evaluar 91 artículos de investigación, se eligieron 53 en función de su relevancia para el tema de estudio.

3. *Presentación del informe de revisión*. La fase final del marco de revisión sistemática consistió en derivar resultados analíticos a partir de las respuestas a las preguntas de investigación. Esto se presenta en la siguiente sección.

3.4. Revisión y Discusión

De acuerdo con la revisión sistemática de la literatura, finalmente se seleccionaron 53 artículos científicos. Las fechas de publicación de los artículos seleccionados abarcan el intervalo de tiempo entre 2014 y 2022. El número de citas de cada una de las publicaciones hasta la fecha se muestra en la Figura 3.7. De estos, el 42 % fueron publicados en las memorias de la Conferencia IEEE sobre Visión por Computadora y Reconocimiento de Patrones (CVPR), la conferencia de referencia en el campo. Además, 17 artículos (32 %) se publicaron en otras memorias de conferencias. Finalmente, los 14 artículos restantes (26 %) fueron publicados en revistas científicas, como se muestra en la Figura 3.8. Además, se encontró que 30 de las 53 publicaciones están disponibles en la web para uso de la comunidad científica en repositorios de GitHub. Cabe mencionar que algunos trabajos comparten su modelo en línea (Xu et al., 2015; Wang and Chan, 2018; Kiros et al., 2014b; Fang et al., 2015; Mao et al., 2015; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Jia et al., 2015; Fu et al., 2015; Johnson et al., 2016; Mao et al., 2016; Yang et al., 2016b; Hendricks et al., 2016; Lu et al., 2017; Chen et al., 2017; Gan et al., 2017b; Tavakoli et al., 2017; Gu et al., 2017; Rennie et al., 2017; Venugopalan et al., 2017; Huang et al., 2019; Cornia et al., 2020; Zhou et al., 2020; Pan et al., 2020; Klein et al., 2022), como se presenta en la Figura 3.9, destacando el uso del lenguaje de programación Python en todos ellos y el uso de los frameworks

TensorFlow y PyTorch. La Figura 3.10 muestra los 200 términos más repetidos encontrados en los 53 artículos científicos. Las palabras “image” y “model” son las más repetidas, pero “attention”, “LSTM” y “sentence” son términos comunes en todos los artículos. También aparecen en la lista los conjuntos de datos y las métricas: “MSCOCO”, “ImageNet”, “BLEU” y “METEOR”. Además, se mencionan nombres de autores, eventos y repositorios en línea: “Karpathy”, “CVPR” y “arXiv”.

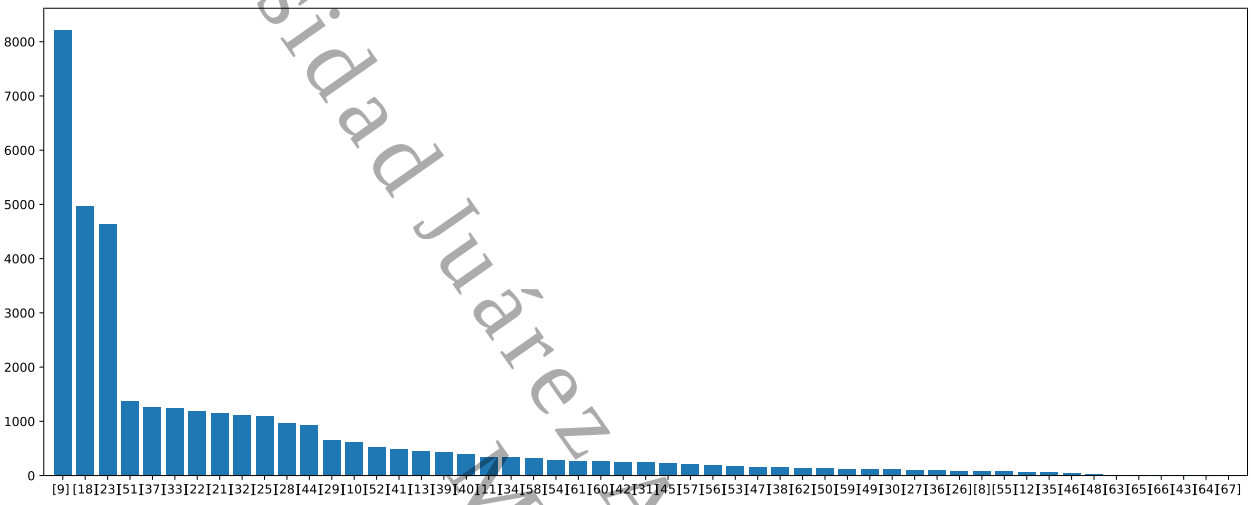

Figura 3.7. Citas por artículo de investigación.

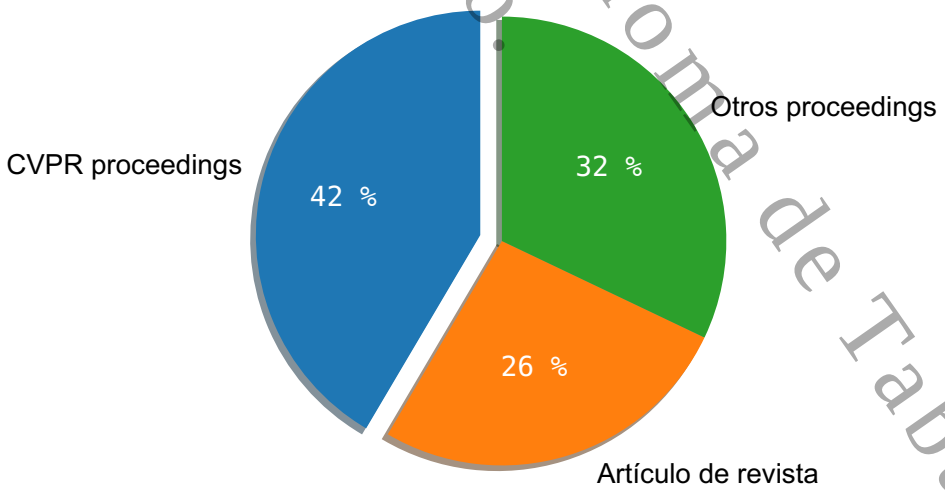

Figura 3.8. Distribución de las fuentes de publicación de los artículos.

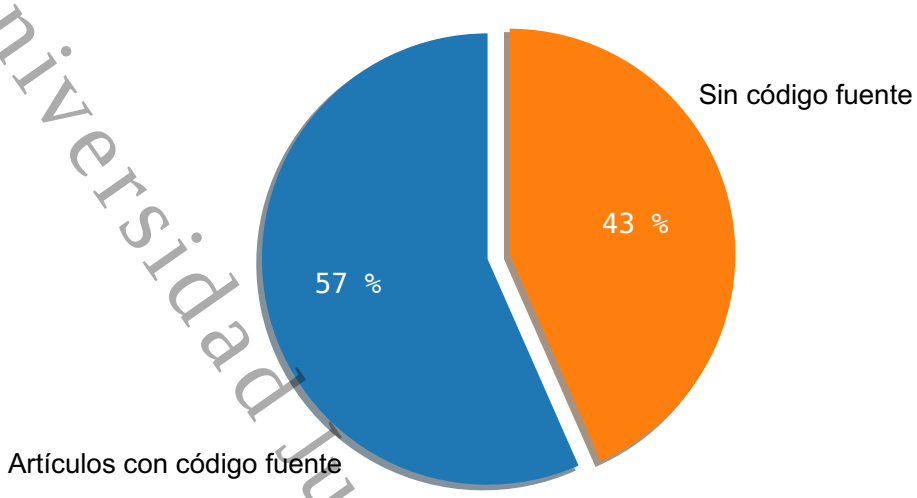

Figura 3.9. Porcentaje de artículos con código fuente en línea.

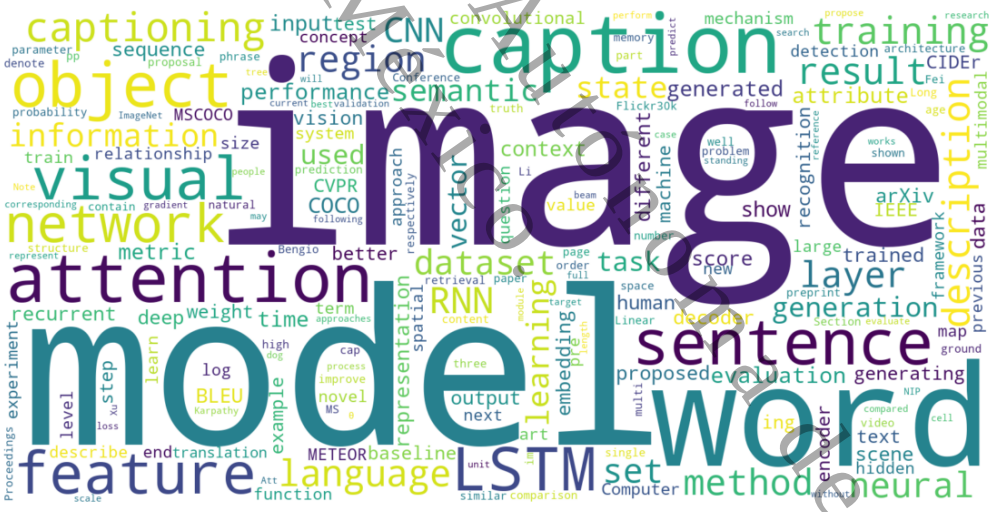

Figura 3.10. Nube de palabras con los términos más representativos encontrados en los artículos relevantes.

Las siguientes subsecciones se introducen a partir de los datos de la Tabla 3.1 y describen las tres arquitecturas que generan descripciones automáticas de imágenes.

Tabla 3.1. Artículos seleccionados sobre descripción automática de imágenes (ordenados por fecha).

Autor, Año	**Arquitectura**	**Conjunto de datos**	**Métricas de evaluación**
Karpathy et al. (2014)	CNN+RNN	Flickr 8K/Flickr 30K	mRank
Mao et al. (2014)	CNN+RNN	Flickr 8K/Flickr 30K, IAPR TC-12	BLEU, mRank
Kiros et al. (2014a)	CNN+RNN	IAPR TC-12, SBU	BLEU, PPLX
Kiros et al. (2014b)	CNN+RNN	Flickr 8K, Flickr 30K	mRank
Chen and Lawrence Zitnick (2015)	CNN+RNN	PASCAL, Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Mao et al. (2015)	CNN+RNN	IAPR TC-12, Flickr 8K/Flickr 30K	BLEU, mRank
Fang et al. (2015)	CNN+RNN	PASCAL, MS COCO	BLEU, METEOR
Karpathy and Fei-Fei (2015)	CNN+RNN	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Vinyals et al. (2015)	CNN+LSTM	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Jia et al. (2015)	CNN+LSTM	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Xu et al. (2015)	CNN+LSTM	Flickr 8K/30K, MS COCO	BLEU, METEOR
Fu et al. (2015)	CNN+LSTM	Flickr 8K/30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Yang et al. (2016b)	CNN+RNN	MS COCO	BLEU, METEOR, CIDEr
Sugano and Bulling (2016)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Mathews et al. (2016)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wang et al. (2016)	CNN+LSTM	Flickr 8K/30K, MS COCO	BLEU, mRank
Johnson et al. (2016)	CNN+LSTM	Visual Genome	METEOR
Mao et al. (2016)	CNN+LSTM	MS COCO	BLEU, METEOR, CIDEr

Tabla 3.1 (continuación)

Autor, Año	Arquitectura	Conjunto de datos	Métricas de evaluación
Tran et al. (2016)	CNN+LSTM	MS COCO, MIT-Adobe FiveK	Evaluación Humana
Ma and Han (2016)	CNN+LSTM	Flickr 8K, UIUC	BLEU, mRank
You et al. (2016)	CNN+LSTM	Flickr 30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Yang et al. (2016a)	CNN+LSTM	Visual Genome	METEOR
Hendricks et al. (2016)	CNN+LSTM	MS COCO, ImageNet	BLEU, METEOR
Yao et al. (2017b)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Lu et al. (2017)	CNN+LSTM	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr
Chen et al. (2017)	CNN+LSTM	Flickr 8K/30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Gan et al. (2017b)	CNN+LSTM	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr
Pedersoli et al. (2017)	CNN+LSTM	MS COCO	BLEU, METEOR, CIDEr
Ren et al. (2017)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wang et al. (2017)	CNN+LSTM	MS COCO, Stock3M	SPICE, METEOR, ROUGE, CIDEr
Tavakoli et al. (2017)	CNN+LSTM	MS COCO, PASCAL	BLEU, METEOR, ROUGE, CIDEr
Liu et al. (2017a)	CNN+LSTM	Flickr 30K, MS COCO	BLEU, METEOR
Gan et al. (2017a)	CNN+LSTM	Flickr 30K	BLEU, METEOR, ROUGE, CIDEr
Liu et al. (2017b)	CNN+LSTM	MS COCO	SPIDEr, Evaluación Humana
Gu et al. (2017)	CNN+LSTM	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr, SPICE
Yao et al. (2017a)	CNN+LSTM	MS COCO, ImageNet	METEOR

Tabla 3.1 (continuación)

Autor, Año	Arquitectura	Conjunto de datos	Métricas de evaluación
Rennie et al. (2017)	CNN+LSTM	MS COCO	BLEU, METEOR, CIDEr, ROUGE
Venugopalan et al. (2017)	CNN+LSTM	MS COCO, ImageNet	METEOR
Zhang et al. (2017)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wu et al. (2018)	CNN+LSTM	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Aneja et al. (2018)	CNN+CNN	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wang and Chan (2018)	CNN+CNN	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Huang et al. (2019)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr, SPICE
Cornia et al. (2020)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr, SPICE
Zhou et al. (2020)	CNN+RNN	MS COCO, Flickr 30K	BLEU, METEOR, CIDEr, SPICE
Ding et al. (2020)	CNN+LSTM	MS COCO, Flickr 30K	BLEU, METEOR, ROUGE, CIDEr
Pan et al. (2020)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr, SPICE
Yang et al. (2021)	CNN+LSTM	MS COCO, Flickr 30K	BLEU, METEOR, ROUGE, CIDEr, SPICE
Zhong and Miyao (2021)	CNN+LSTM	MS COCO, Flickr 30K	BLEU, METEOR, ROUGE, CIDEr, SPICE
Tian et al. (2021)	CNN+LSTM	MS COCO, Flickr 30K	BLEU, METEOR, ROUGE, CIDEr, SPICE
Klein et al. (2022)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr

Tabla 3.1 (continuación)

Autor, Año	Arquitectura	Conjunto de datos	Métricas de evaluación
Deng et al. (2022)	CNN+LSTM	Visual Genome	mRank
Fei (2022)	CNN+LSTM	MS COCO	BLEU, METEOR, ROUGE, CIDEr, SPICE

3.4.1. Principales Arquitecturas

- CNN + RNN. En esta arquitectura, se utiliza una CNN para extraer las características de la imagen, mientras que una RNN se emplea para generar la descripción. Un total de 10 trabajos de los 53 (19 %) siguen este método. Cabe destacar que esta arquitectura se emplea en los primeros trabajos sobre descripción automática de imágenes utilizando el enfoque encoder-decoder.
- CNN + LSTM. Esta arquitectura utiliza una CNN como codificador y una RNN con módulos LSTM para evitar el problema del desvanecimiento del gradiente. La mayoría de los trabajos, incluidos los más recientes (41 de 53, lo que representa el 77 %), siguen este método.
- CNN + CNN. Esta arquitectura utiliza dos CNN; la primera se encarga de extraer las características de la imagen, y la segunda de generar la descripción de la imagen a partir de los resultados de la primera CNN. Solo dos trabajos (4 %) utilizan este método.

La Figura 3.11 muestra la distribución de los enfoques encoder-decoder encontrados en la literatura.

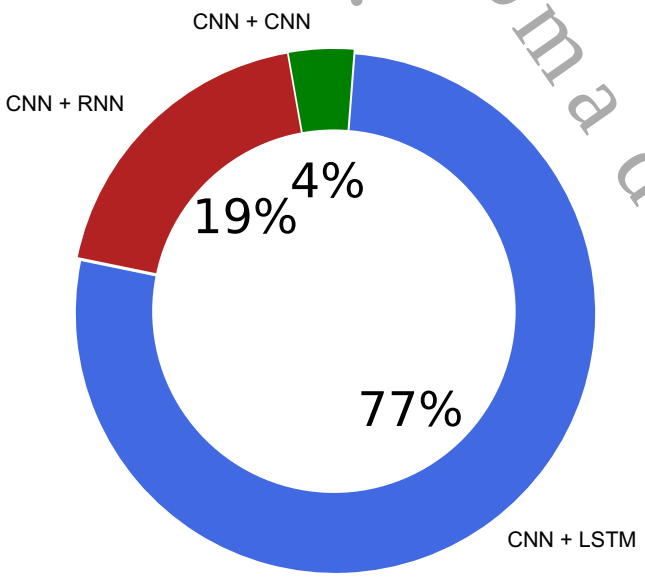

Figura 3.11. Distribución de las arquitecturas encoder-decoder.

De acuerdo con la revisión de la literatura, se observa que una CNN se implementa de manera consistente como módulo codificador. Por otro lado, el módulo decodificador puede implementarse utilizando tres arquitecturas diferentes: una RNN simple, LSTM u otra CNN. La mayoría de los modelos de descripción automática de imágenes utilizan una arquitectura LSTM en el lado del decodificador, debido a su efectividad para memorizar secuencias de datos mediante las celdas de memoria.

3.4.2. Arquitectura CNN + RNN

El año 2014 puede considerarse como el punto de partida para la descripción automática de imágenes utilizando aprendizaje profundo. Karpathy et al. (2014) popularizaron el uso del enfoque encoder-decoder. La novedad de este enfoque fue que era posible mapear las imágenes a un conjunto fijo de oraciones utilizando técnicas de aprendizaje profundo.

Siguiendo este enfoque, Mao et al. (2014) presentaron un modelo RNN multimodal (m-RNN) para generar descripciones novedosas en forma de oraciones que explicaran el contenido de las imágenes. Posteriormente, Mao et al. (2015) mejoraron su trabajo anterior utilizando representaciones de imagen más complejas y modelos de lenguaje más sofisticados.

Kiros et al. (2014a) mapearon las características de la imagen en un espacio compartido con las características de las palabras mediante una técnica de espacio multimodal. Además, mejoraron su enfoque en Kiros et al. (2014b) proponiendo un nuevo modelo de lenguaje neuronal llamado modelo de lenguaje neuronal estructura-contenido (SC-NLM). Este modelo permitió una mejor extracción de las estructuras de las oraciones, mejorando así la generación de subtítulos.

Chen and Lawrence Zitnick (2015) propusieron un modelo de proyección inversa. Este modelo incluye una capa recurrente adicional que realiza una proyección inversa, lo que permite una actualización dinámica de las representaciones visuales de una imagen a partir de las palabras generadas. Fang et al. (2015) trabajaron con subregiones de la imagen en lugar de utilizar la imagen completa; emplearon AlexNet y VGG16Net para extraer características de dichas subregiones.

Karpathy and Fei-Fei (2015) continuaron con su trabajo seminal proponiendo un modelo que utiliza regiones específicas de una imagen para generar descripciones en lenguaje natural. Su modelo utilizó una combinación novedosa de CNN y RNN denominada alineaciones visuales-semánticas profundas.

Posteriormente, Yang et al. (2016b) propusieron un sistema de generación de subtítulos para imágenes que explota las estructuras paralelas entre imágenes y oraciones utilizando una RNN clásica.

Más recientemente, Zhou et al. (2020) introdujeron un método que utiliza una red de transformadores multicapa compartida, la cual también puede ajustarse para tareas de generación de lenguaje y visión.

3.4.3. Arquitectura CNN + LSTM

Vinyals et al. (2015) introdujeron la implementación de una arquitectura LSTM para el módulo decodificador. La red codificadora es una CNN, y la red decodificadora es un conjunto de capas LSTM. En cinco trabajos (Liu et al. (2017b); Gan et al. (2017a); Gu et al. (2017); Yao et al. (2017a); Wang et al. (2016)), los autores utilizan LSTM en el lado del decodificador siguiendo este mismo enfoque.

Los siguientes cuatro trabajos presentan modelos que utilizan los objetos y las relaciones dentro de la escena para generar las descripciones. Mao et al. (2016) propusieron una expresión de referencia, que genera una descripción de un objeto o región específica. Este método permite inferir el objeto o región que se está describiendo, generando así oraciones relativamente no ambiguas. Hendricks et al. (2016) propusieron un método llamado ”deep compositional captioner” (DCC), que puede generar descripciones de objetos que no están presentes en los conjuntos de datos de imágenes con oraciones emparejadas. Yao et al. (2017b) propusieron un método que utiliza un mecanismo de copiado y un conjunto de datos separado de reconocimiento de objetos, logrando así generar descripciones de objetos nuevos no encontrados en las oraciones de entrenamiento. Wang et al. (2017) propusieron “skeleton key”, que primero localiza los objetos y sus interacciones, y luego identifica y extrae los atributos relevantes para generar las descripciones de imágenes. El método descompone la descripción en dos partes: oraciones esqueleto y oraciones con atributos.

Los dos trabajos siguientes enriquecen la descripción utilizando información de los rostros que aparecen en la escena. Sugano and Bulling (2016) usaron la información de la dirección de la mirada presentada en los rostros de las imágenes para enriquecer la descripción de la escena. Adicionalmente, Tran et al. (2016) introdujeron un método capaz de detectar un conjunto diverso de conceptos visuales y generar oraciones que reconocen celebridades; este método alcanzó un rendimiento notable.

Los siguientes siete trabajos emplean diversas técnicas para mejorar la descripción de imágenes. Mathews et al. (2016) propusieron un método llamado SentiCap, que genera descripciones de imágenes con sentimientos negativos o positivos. Wu et al. (2018) propusieron un método para lograr descripciones de alto nivel implementando un esquema de selección de conocimiento guiado por preguntas para descartar información irrelevante, logrando una mejor descripción de las imágenes. Yang et al. (2021) propusieron un nuevo modelo, CaptionNet, para ayudar a la LSTM a evitar la acumulación de errores derivados de palabras irrelevantes durante la generación de subtítulos de imágenes. Zhong and Miyao (2021) propusieron un marco para que los subtítulos de imágenes incluyan palabras específicas y tengan una mejor estructura sintáctica. También propusieron un modelo consciente de la estructura de dependencia sintáctica (SD-SAM) para apoyar dicho marco. Cornia et al. (2020) propusieron un método llamado M2, que incluye un transformador con malla de memoria para la generación de oraciones.

Deng et al. (2022) propusieron un novedoso marco de aprendizaje jerárquico de memoria (HML) para entrenar con oraciones que contienen predicados gruesos y oraciones con predicados delgados. Esto permite generar oraciones más detalladas en sus descripciones de la escena que los enfoques tradicionales. Más recientemente, Fei (2022) propuso un modelo que genera descripciones que explotan eficazmente el contexto global de la escena sin implicar un costo adicional de inferencia. El modelo se entrena con dos conjuntos: uno que contiene las etiquetas de las descripciones y otro que incluye la descripción del contexto general de la imagen.

En las siguientes secciones, agrupamos los trabajos restantes con arquitectura CNN + LSTM en tres enfoques: descripción de imágenes basada en atención, basada en semántica y basada en aprendizaje por refuerzo.

3.4.3.1. Descripción de Imágenes Basada en Atención

Los modelos que utilizan mecanismos de atención se enfocan en las regiones prominentes de una imagen para generar oraciones, considerando la escena en su totalidad.

Xu et al. (2015) fueron los primeros en introducir un método de descripción de imágenes basado en atención. Los autores desarrollaron una codificación más rica que permite al decodificador aprender en qué parte de la imagen debe enfocar su atención mientras genera cada palabra de la descripción. Actualmente, este es el enfoque más exitoso para la generación automática de subtítulos de imágenes, y también es el artículo más citado de la colección, con más de 8000 citas hasta la fecha (Figura 3.7). Siguiendo esta línea de descripción basada en atención, Fu et al. (2015) propusieron un método capaz de extraer información de la escena en función de la relación semántica entre la información textual y visual. Johnson et al. (2016) propusieron DenseCap, un método que localiza las regiones salientes de una imagen y luego genera descripciones para cada una de ellas. Lu et al. (2017) propusieron un novedoso modelo de atención adaptativa con un ”centinela visual”; este centinela ayuda a predecir palabras no visuales como “el” y “de”. Chen et al. Chen et al. (2017) introdujeron una nueva CNN denominada SCA-CNN que incorpora atención espacial y por canales en una CNN.

Pedersoli et al. (2017) propusieron un método que utiliza mecanismos de atención para asociar regiones de una imagen con palabras de los subtítulos generados por el decodificador. Tavakoli et al. (2017) propusieron un método de descripción de imágenes basado en atención. Este método se basa en cómo los humanos tienden a describir primero los objetos más importantes antes que los menos relevantes.

Huang et al. (2019) propusieron un método en el que utilizan un módulo denominado “attention on attention” (AoA), que les permite determinar la relevancia entre los resultados de la atención y la consulta. Además, aplicaron el módulo (AoA) tanto al codificador como al decodificador de su modelo de descripción.

Pan et al. Pan et al. (2020) introdujeron un bloque de atención unificado denominado X-Linear, que permite a la red realizar razonamiento multimodal para aprovechar selectivamente la información visual. Ding et al. (2020) propusieron un modelo que utiliza dos mecanismos de atención: uno impulsado por estímulos y otro impulsado por conceptos. Introdujeron la teoría de la atención desde la psicología para la generación de subtítulos de imágenes, logrando un buen desempeño. Klein et al. (2022) presentaron un marco de auto-codificador variacional que permite aprovechar las características de las regiones de la imagen mediante un mecanismo de atención para generar subtítulos coherentes. Sus experimentos demuestran que este enfoque genera subtítulos precisos y diversos, con estilos variados expresados en la imagen.

3.4.3.2. Descripción de Imágenes Basada en Semántica

Las descripciones de imágenes basadas en semántica tienen como objetivo enriquecer el lenguaje para generar oraciones con conceptos semánticos.

Jia et al. (2015) propusieron una modificación a la arquitectura LSTM, denominada LSTM guiada (gLSTM), que permite la generación de oraciones largas. En esta arquitectura, la red puede extraer información semántica de cada oración añadiéndola a cada compuerta y al estado de la celda LSTM. Ma and Han (2016) propusieron un método que genera descripciones semánticas significativas utilizando palabras estructurales en la siguiente forma: (objeto, atributo, actividad y escena). Yang et al. (2016b) propusieron un método de subtitulación densa. Este método consiste en utilizar las características visuales de una región y los subtítulos proporcionados para dicha región, combinándolos con las características contextuales y aplicando un mecanismo de inferencia para lograr una descripción enriquecida semánticamente. Gan et al. (2017b) desarrollaron una red de composición semántica (SCN) para la generación de subtítulos de imágenes, en la cual los conceptos semánticos se detectan desde la imagen para alimentar la red LSTM. Venugopalan et al. (2017) propusieron un método que utiliza fuentes externas, como conocimiento semántico extraído de texto no anotado e imágenes etiquetadas de conjuntos de datos de reconocimiento de objetos. Tian et al. (2021) propusieron una red de información de contexto semántico multinivel (MSCI). El modelo actualiza las diferentes características semánticas de una imagen y luego implementa una red de extracción de información contextual para aprovechar la información entre las capas semánticas, mejorando así la precisión en la generación de tareas visuales.

Cabe señalar que dos trabajos presentaron modelos que combinan mecanismos basados en atención y en semántica. You et al. (2016) introdujeron un modelo de atención semántica que permite enfocar selectivamente en los atributos semánticos de una imagen. Liu et al. (2017a) propusieron un método que utiliza dos tipos de modelos de atención supervisada: supervisión fuerte con anotación de alineación y supervi-

sión débil con etiquetado semántico. Esto permite corregir el mapa de atención en cada paso temporal.

3.4.3.3. Descripción de Imágenes Basada en Aprendizaje por Refuerzo

El aprendizaje por refuerzo es un enfoque de aprendizaje automático en el que un agente busca descubrir datos y etiquetas mediante la exploración y una señal de recompensa.

Ren et al. (2017) introdujeron un método novedoso para describir imágenes basado en aprendizaje por refuerzo, el cual predice la siguiente mejor palabra en la oración con la ayuda de dos redes: una red política y una red de valor. Zhang et al. (2017) presentaron un método basado en el aprendizaje por refuerzo actor–crítico, que propone una ventaja por token al utilizar LSTM, logrando así descripciones de mejor calidad. Rennie et al. (2017) propusieron un método de descripción de imágenes basado en aprendizaje por refuerzo que genera descripciones altamente efectivas mediante la implementación de un algoritmo de inferencia temporal.

3.4.4. Arquitectura CNN + CNN

Aneja et al. (2018) propusieron una arquitectura que utiliza únicamente redes convolucionales tanto en el codificador como en el decodificador (no utilizaron ninguna función recursiva). Siguiendo este mismo enfoque, Wang and Chan (2018) también utilizaron dos CNN; el modelo resultante fue tres veces más rápido que el modelo “show-and-tell” (que emplea LSTM como módulo decodificador). Ambos trabajos emplearon una CNN tanto en el codificador como en el decodificador.

3.4.5. Conjuntos de Datos

La mayoría de los conjuntos de datos utilizados para entrenar modelos de descripción automática de imágenes también se emplean en tareas relacionadas, como la detección de rostros. Por esta razón, algunos de estos conjuntos también contienen clases y cuadros delimitadores (bounding boxes) en sus imágenes. Los conjuntos de datos encontrados en la literatura fueron:

- **MS COCO** (Chen et al., 2015). El conjunto de datos Microsoft Common Objects in Context (COCO) fue desarrollado por el equipo de Microsoft y está orientado a la comprensión de escenas. Contiene imágenes de escenas complejas de la vida diaria y puede utilizarse en tareas como reconocimiento, segmentación y descripción de imágenes. El conjunto incluye 165,482 imágenes y un archivo de texto con casi un millón de descripciones. Es el conjunto de datos más utilizado, apareciendo en 77 % de los trabajos revisados (41 de 53).

- **Flickr8k**/**Flickr30k** (Hodosh et al., 2013; Plummer et al., 2015). Las imágenes del conjunto Flickr8k provienen del sitio web de álbumes de fotos Flickr, de Yahoo, y contiene 8000 imágenes. Flickr30k es una versión extendida con 31,783 imágenes recolectadas del mismo sitio. Generalmente representan escenas del mundo real y cada imagen tiene cinco descripciones. Flickr8k se usa en 13 trabajos (25 %) y Flickr30k en 11 (21 %).
- **Visual Genome** (Krishna et al., 2017). Este conjunto fue creado para fomentar investigaciones que conecten conceptos estructurados en imágenes con el lenguaje. Contiene 108,077 imágenes y 5.4 millones de descripciones de regiones. Es utilizado en tres trabajos (Johnson et al., 2016; Yang et al., 2016b; Deng et al., 2022), lo que representa un 6 %.
- **IAPR–TC12** (Grubinger et al., 2006). Contiene 20,000 imágenes recopiladas de diversas fuentes, como deportes, personas, animales y paisajes. Las imágenes están acompañadas de subtítulos en varios idiomas. Es utilizado en tres trabajos (Kiros et al., 2014a; Mao et al., 2014, 2015), es decir, un 6 %.
- **Stock3M** (Wang et al., 2017). Este conjunto contiene más de 3.2 millones de imágenes subidas por usuarios y es 26 veces más grande que MS COCO. Las imágenes son muy variadas e incluyen personas, naturaleza y objetos artificiales. Solo un trabajo lo utiliza (Wang et al., 2017) (2 %).
- **MIT-Adobe FiveK** (Bychkovsky et al., 2011). Contiene 5000 imágenes que abarcan diversas escenas, sujetos y condiciones de iluminación. Está compuesto principalmente por imágenes de personas, naturaleza y objetos fabricados por humanos. Solo un trabajo lo utiliza (Tran et al., 2016) (2 %).
- **SBU Captions** (Ordonez et al., 2011). Es un conjunto antiguo que contiene imágenes acompañadas de descripciones breves. Se utiliza para inducir embeddings de palabras aprendidos a partir de texto e imágenes. Contiene un millón de imágenes con subtítulos visualmente relevantes. Solo un trabajo lo utiliza (Kiros et al., 2014a) (2 %).
- **PASCAL** (Everingham et al., 2014). Proporciona un conjunto estandarizado de imágenes para el reconocimiento de clases de objetos y herramientas comunes para acceder a las anotaciones. El conjunto de entrenamiento y validación tiene 10,103 imágenes, y el de prueba 9,637 imágenes. Es utilizado en tres trabajos (Tran et al., 2016; Fang et al., 2015; Tavakoli et al., 2017) (6 %).
- **UIUC** (Li and Fei-Fei, 2007). Contiene categorías de ocho eventos deportivos: remo, bádminton, polo, bochas, snowboard, croquet, navegación y escalada. Las imágenes están divididas según el nivel de dificultad. Solo un trabajo lo utiliza (Ma and Han, 2016) (2 %).

- **ImageNet** (Deng et al., 2009). Consiste en 150,000 fotografías etiquetadas manualmente, recolectadas de Flickr y otros motores de búsqueda. Es utilizado en tres trabajos (Hendricks et al., 2016; Rennie et al., 2017; Venugopalan et al., 2017) (6 %).

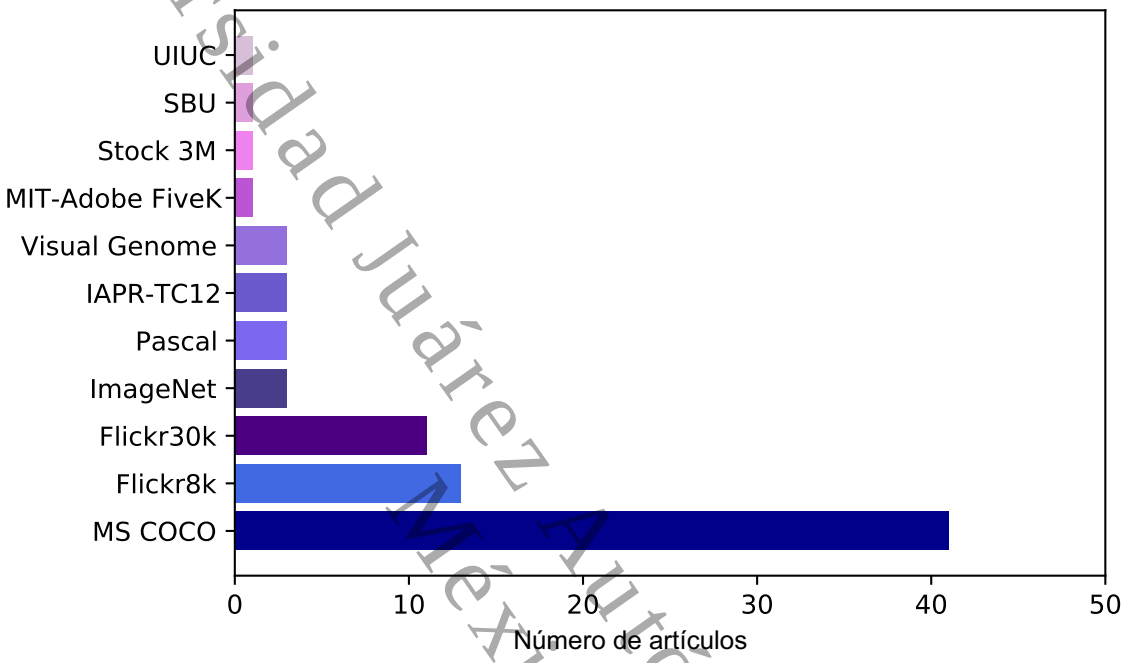

Figura 3.12. Número de trabajos que utilizan cada conjunto de datos.

x

3.4.6. Métricas de Evaluación

Al evaluar un modelo basado en la calidad del lenguaje generado, es necesario utilizar métricas específicas, ya que las métricas tradicionales como precisión, exactitud o sensibilidad no pueden aplicarse directamente al comparar dos textos en lenguaje natural. Por esta razón, en la descripción de imágenes se han utilizado una serie de estándares, originalmente provenientes de la traducción automática, para comparar las descripciones. Las métricas de evaluación encontradas en la literatura fueron:

- **BLEU (Bilingual evaluation understudy)** (Papineni et al., 2002). Es la métrica más utilizada en la práctica. Su propósito original no era la descripción de imágenes, sino la traducción automática. Se basa en la evaluación de la tasa de coincidencia de n-gramas (secuencias continuas de palabras) entre la traducción generada y la oración de referencia. Un 100 % de los artículos revisados usaron esta métrica.

$$\mathrm{BP} = \begin{cases} 1 & \text{si } c > r \\ e^{(1-r/c)} & \text{si } c \leq r \end{cases} \tag{3.1}$$

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3.2}$$

- **ROUGE (Recall-oriented understudy for gisting evaluation)** (Lin, 2004). Conjunto de métricas comúnmente usadas para evaluar resúmenes automáticos y traducción. Se basa en la comparación de n-gramas entre la hipótesis y las referencias. Fue usada en 33 de 53 trabajos (62 %).

$$\text{ROUGE } -L = \frac{\left(1+\beta^2\right) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{3.3}$$

- **METEOR (Metric for evaluation of translation with explicit ordering)** (Lavie and Agarwal, 2007). Métrica para traducción automática que realiza una alineación entre la oración generada y la oración de referencia. Fue usada en 47 trabajos (89 %).

$$F_{\text{mean}} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \tag{3.4}$$

$$Pen = \gamma \cdot frag^{\beta} \tag{3.5}$$

$$\text{score } = (1 - \text{ Pen }) \cdot F_{\text{mean}} \tag{3.6}$$

- **CIDEr (Consensus-based image description evaluation)** (Vedantam et al., 2015). Métrica diseñada para evaluar descripciones de imágenes. Usada en 15 trabajos (25 %).

$$\mathrm{CIDEr}_n\left(c_i, S_i\right) = \frac{1}{m} \sum_{j} \frac{\boldsymbol{g^n}\left(c_i\right) \cdot \boldsymbol{g^n}\left(s_{ij}\right)}{\left\|\boldsymbol{g}^n\left(c_i\right)\right\| \left\|\boldsymbol{g^n}\left(s_{ij}\right)\right\|} \tag{3.7}$$

- **SPICE** (Anderson et al., 2016). Mide la similitud semántica entre descripciones y objetos de la imagen. Usada en 24 trabajos (45 %).

$$G(c) = \langle O(c), E(c), K(c) \rangle \tag{3.8}$$

$$T(G(c)) \triangleq O(c) \cup E(c) \cup K(c) \tag{3.9}$$

$$P(c,S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|} \tag{3.10}$$

$$R(c,S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|} \tag{3.11}$$

$$\text{SPICE}(c,S) = F_1(c,S) = \frac{2 \cdot P(c,S) \cdot R(c,S)}{P(c,S) + R(c,S)} \tag{3.12}$$

- **SPIDEr** Liu et al. (2017b). Combina CIDEr y SPICE mediante optimización por gradiente de política. Usada en 34 trabajos (64 %).

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} V_\theta \left(s_0 \mid \mathbf{x}^n, \mathbf{y}^n\right) \tag{3.13}$$

- **mRank (Matrix rank)** (Socher et al., 2014). Mide el rango medio de la descripción correcta por imagen. Usada en 5 trabajos (9 %).

$$\text{IoU} = \frac{\text{area } (B_p \cap B_{gt})}{\text{area } (B_p \cup B_{gt})} \tag{3.14}$$

$$\text{recall } = \frac{TP}{TP + FN} \tag{3.15}$$

$$\text{AR } = 2 \int_{0.5}^{1} \text{recall}(o) do \tag{3.16}$$

$$\text{mAR } = \frac{\sum_{i=1}^{K} AR_i}{K} \tag{3.17}$$

- **PPLX (Perplexity)** (Kiros et al., 2014a). Propuesta por Kiros et al. para evaluar el uso de embeddings preentrenados. Solo un trabajo la utilizó.

$$\log_2 \mathcal{C}\left(w_{1:n} \mid \mathbf{x}\right) = -\frac{1}{N} \sum_{w_{1:n}} \log_2 P\left(w_n = i \mid w_{1:n-1}, \mathbf{x}\right) \tag{3.18}$$

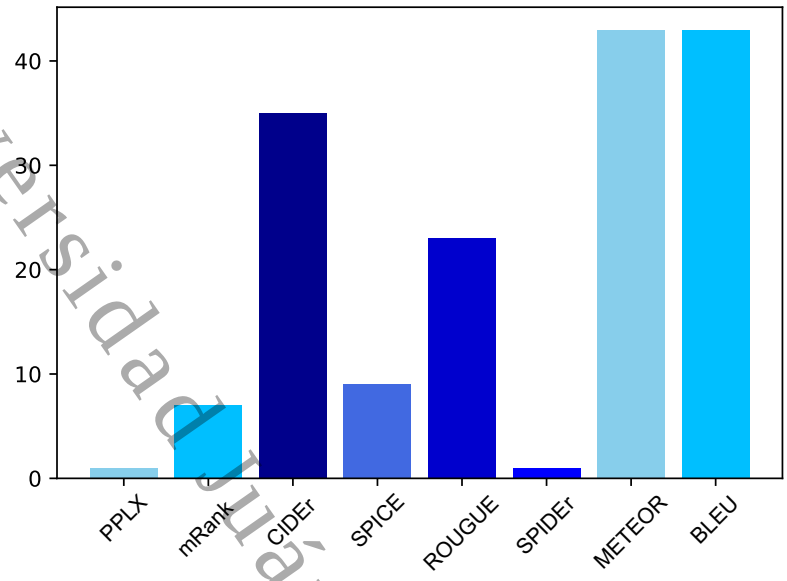

Figura 3.13. Uso de métricas de evaluación entre los 53 artículos revisados.

3.5. Conclusiones y Direcciones Futuras

En este artículo, revisamos y analizamos estudios sobre la descripción de imágenes, enfocándonos en arquitecturas encoder-decoder. Tras analizar 53 artículos de investigación, concluimos que:

- La arquitectura predominante para la descripción automática de imágenes emplea una red neuronal convolucional (CNN) como codificador y una red LSTM como decodificador.
- El conjunto de datos más utilizado para entrenar y evaluar los modelos es MS COCO, empleado por casi todos los trabajos revisados.
- Todos los artículos revisados utilizan más de una métrica para comparar el rendimiento de los modelos propuestos, destacándose BLEU y METEOR como las más utilizadas.

Con base en nuestro estudio, algunas direcciones de investigación para trabajos futuros sobre la descripción automática de imágenes se enfocarán en los siguientes aspectos:

- **Modelos multilingües:** Los modelos y avances actuales en la generación automática de descripciones de imágenes se han centrado exclusivamente en el idioma inglés. Sería interesante estudiar otros idiomas o conjuntos de datos multilingües.
- **Cantidad de datos para el entrenamiento:** La mayoría de los modelos actuales utilizan un enfoque de aprendizaje supervisado, por lo que requieren una gran cantidad de datos etiquetados. Por esta

razón, el aprendizaje semi-supervisado, no supervisado y por refuerzo tendrán una mayor presencia en la creación de modelos futuros para la descripción automática de imágenes.

- **Variedad de conjuntos de datos:** La precisión de las descripciones generadas por los modelos existentes depende del conjunto de datos utilizado, y actualmente hay pocos disponibles. Sería interesante contar con más conjuntos de datos, cada vez más diversos, para futuras investigaciones en este campo.

Bibliografía

Amirian, S., Rasheed, K., Taha, T. R., and Arabnia, H. R. (2020). Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap. *IEEE Access*, 8:218386–218400. https://doi.org/10.1109/ACCESS.2020.3042484.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-46454-1_24.

Aneja, J., Deshpande, A., and Schwing, A. G. (2018). Convolutional image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2018/papers/Aneja_Convolutional_Image_Captioning_CVPR_2018_paper.pdf.

Bychkovsky, V., Paris, S., Chan, E., and Durand, F. (2011). Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. https://doi.org/10.1109/CVPR.2011.5995413.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., and Chua, T.-S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Chen_SCA-CNN_Spatial_and_CVPR_2017_paper.pdf.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*. https://www.cs.cmu.edu/~jeanoh/16-785/papers/chen-arxiv2015-mscoco-metrics.pdf.

Chen, X. and Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption

generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2015.7298856.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Cornia_Meshed-Memory_Transformer_for_Image_Captioning_CVPR_2020_paper.pdf.

Dai, B., Fidler, S., Urtasun, R., and Lin, D. (2017). Towards diverse and natural image descriptions via a conditional gan. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Dai_Towards_Diverse_and_ICCV_2017_paper.pdf.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Deng, Y., Li, Y., Zhang, Y., Xiang, X., Wang, J., Chen, J., and Ma, J. (2022). Hierarchical memory learning for fine-grained scene graph generation. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *Computer Vision – ECCV 2022*, pages 266–283, Cham. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19812-0_16.

Ding, S., Qu, S., Xi, Y., and Wan, S. (2020). Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*, 398:520–530. https://doi.org/10.1016/j.neucom.2019.04.095.

Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A. (2014). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136. https://doi.org/10.1007/s11263-014-0733-5.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Lawrence Zitnick, C., and Zweig, G. (2015). From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2015/papers/Fang_From_Captions_to_2015_CVPR_paper.pdf.

Fei, Z. (2022). Efficient modeling of future context for image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM. https://feizc.github.io/resume/future.pdf.

Fu, K., Jin, J., Cui, R., Sha, F., and Zhang, C. (2015). Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2321–2334. https://doi.org/10.1109/TPAMI.2016.2642953.

Gan, C., Gan, Z., He, X., Gao, J., and Deng, L. (2017a). Stylenet: Generating attractive visual captions with styles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Gan_StyleNet_Generating_Attractive_CVPR_2017_paper.pdf.

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., and Deng, L. (2017b). Semantic compositional networks for visual captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Gan_Semantic_Compositional_Networks_CVPR_2017_paper.pdf.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The iapr tc12 benchmark: A new evaluation resource for visual information systems. *Workshop Ontoimage*, 2. http://www.thomas.deselaers.de/publications/papers/grubinger_lrec06.pdf.

Gu, J., Wang, G., Cai, J., and Chen, T. (2017). An empirical study of language cnn for image captioning. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Gu_An_Empirical_Study_ICCV_2017_paper.pdf.

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., and Darrell, T. (2016). Deep compositional captioning: Describing novel object categories without paired training data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2016/papers/Hendricks_Deep_Compositional_Captioning_CVPR_2016_paper.pdf.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899. https://www.ijcai.org/Proceedings/15/Papers/593.pdf.

Houdt, G. V., Mosquera, C., and Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8):5929–5955. https://doi.org/10.1007/s10462-020-09838-1.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). Attention on attention for image captioning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2019/papers/Huang_Attention_on_Attention_for_Image_Captioning_ICCV_2019_paper.pdf.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_iccv_2015/papers/Jia_Guiding_the_Long-Short_ICCV_2015_paper.pdf.

Jiang, W., Li, X., Hu, H., Lu, Q., and Liu, B. (2021). Multi-gate attention network for image captioning. *IEEE Access*, 9:69700–69709. https://doi.org/10.1109/ACCESS.2021.3067607.

Johnson, J., Karpathy, A., and Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2016/papers/Johnson_DenseCap_Fully_Convolutional_CVPR_2016_paper.pdf.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://cs.stanford.edu/people/karpathy/cvpr2015.pdf.

Karpathy, A., Joulin, A., and Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc. https://cs.stanford.edu/people/karpathy/nips2014.pdf.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014a). Multimodal neural language models. In Xing, E. P. and Jebara, T., editors, *31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 595–603, Bejing, China. PMLR. http://proceedings.mlr.press/v32/kiros14.pdf.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*. https://arxiv.org/pdf/1411.2539.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical report, Keele, UK, Keele University. https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf.

Klein, F., Mahajan, S., and Roth, S. (2022). Diverse image captioning with grounded style. In *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings*, pages 421–436. Springer. https://doi.org/10.1007/978-3-030-92659-5_27.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73. https://doi.org/10.1007/s11263-016-0981-7.

Lavie, A. and Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics. https://aclanthology.org/W05-0909.pdf.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. https://doi.org/10.1038/nature14539.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404. https://dl.acm.org/doi/10.5555/109230.109279.

Li, L.-J. and Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. https://doi.org/10.1109/ICCV.2007.4408872.

Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, pages 74–81. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/was2004.pdf.

Liu, C., Mao, J., Sha, F., and Yuille, A. (2017a). Attention correctness in neural image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). https://arxiv.org/pdf/1605.09553.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., and Murphy, K. (2017b). Improved image captioning via policy gradient optimization of spider. In *IEEE International Conference on Computer Vision (ICCV)*. https://doi.org//10.1109/ICCV.2017.100.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Lu_Knowing_When_to_CVPR_2017_paper.pdf.

Ma, S. and Han, Y. (2016). Describing images by feeding lstm with structural words. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. https://doi.org//10.1109/ICME.2016.7552883.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. coursera.org/learn/information-technology-it-fundamentals-for-everyone/home/week/1.

Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*. https://arxiv.org/pdf/1410.1090.

Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. L. (2015). Deep captioning with multimodal recurrent neural networks (m-rnn). In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. https://www.cs.jhu.edu/~ayuille/Pubs15/JunhuaMaoDeepICLR2015.pdf.

Mathews, A., Xie, L., and He, X. (2016). SentiCap: Generating image descriptions with sentiments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1). https://doi.org//10.1609/aaai.v30i1.10475.

Mohamad Nezami, O., Dras, M., Wan, S., Paris, C., and Hamey, L. (2019). Towards generating stylized image captions via adversarial training. In Nayak, A. C. and Sharma, A., editors, *PRICAI 2019: Trends in Artificial Intelligence*, pages 270–284, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-29908-8_22.

Ordonez, V., Kulkarni, G., and Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc. http://www.tamaraberg.com/papers/generation_nips2011.pdf.

Pan, Y., Yao, T., Li, Y., and Mei, T. (2020). X-linear attention networks for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Pan_X-Linear_Attention_Networks_for_Image_Captioning_CVPR_2020_paper.pdf.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014). How to construct deep recurrent neural networks. In *Second International Conference on Learning Representations (ICLR 2014)*. https://arxiv.org/pdf/1312.6026.

Pedersoli, M., Lucas, T., Schmid, C., and Verbeek, J. (2017). Areas of attention for image captioning. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Pedersoli_Areas_of_Attention_ICCV_2017_paper.pdf.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_iccv_2015/papers/Plummer_Flickr30k_Entities_Collecting_ICCV_2015_paper.pdf.

Ren, Z., Wang, X., Zhang, N., Lv, X., and Li, L.-J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Ren_Deep_Reinforcement_Learning-Based_CVPR_2017_paper.pdf.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Rennie_Self-Critical_Sequence_Training_CVPR_2017_paper.pdf.

Sarkar, D., Bali, R., and Sharma, T. (2018). *Practical Machine Learning with Python*. Apress. https://doi.org/10.1007/978-1-4842-3207-1.

Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., and Schiele, B. (2017). Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Shetty_Speaking_the_Same_ICCV_2017_paper.pdf.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218. https://aclanthology.org/Q14-1017.pdf.

Sugano, Y. and Bulling, A. (2016). Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*. https://arxiv.org/pdf/1608.05203.

Tavakoli, H. R., Shetty, R., Borji, A., and Laaksonen, J. (2017). Paying attention to descriptions generated by image captioning models. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Tavakoli_Paying_Attention_to_ICCV_2017_paper.pdf.

The American Anthropological Association (2019). Guidelines for creating image. The American Anthropological Association. https://americananthro.org/.

Tian, P., Mo, H., and Jiang, L. (2021). Image caption generation using multi-level semantic context information. *Symmetry*, 13(7). https://doi.org/10.3390/sym13071184.

Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., and Sienkiewicz, C. (2016). Rich image captioning in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. https://openaccess.thecvf.com/content_cvpr_2016_workshops/w12/papers/Tran_Rich_Image_Captioning_CVPR_2016_paper.pdf.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.pdf.

Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., and Saenko, K. (2017). Captioning images with diverse objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Venugopalan_Captioning_Images_With_CVPR_2017_paper.pdf.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vinyals_Show_and_Tell_2015_CVPR_paper.pdf.

Wang, H., Qin, Z., and Wan, T. (2018). Text generation based on generative adversarial nets with latent variables. In Phung, D., Tseng, V. S., Webb, G. I., Ho, B., Ganji, M., and Rashidi, L., editors, *Advances in Knowledge Discovery and Data Mining*, pages 92–103, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-93037-4_8.

Wang, M., Song, L., Yang, X., and Luo, C. (2016). A parallel-fusion rnn-lstm architecture for image caption generation. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4448–4452. https://doi.org//10.1109/ICIP.2016.7533201.

Wang, Q. and Chan, A. B. (2018). Cnn+ cnn: Convolutional decoders for image captioning. *arXiv preprint arXiv:1805.09019*. https://arxiv.org/pdf/1805.09019.

Wang, Y., Lin, Z., Shen, X., Cohen, S., and Cottrell, G. W. (2017). Skeleton key: Image captioning by skeleton-attribute decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_Skeleton_Key_Image_CVPR_2017_paper.pdf.

Wu, Q., Shen, C., Wang, P., Dick, A., and Hengel, A. v. d. (2018). Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1367–1381. https://doi.org/10.1109/TPAMI.2017.2708709.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning (ICML)*, pages 2048–2057. https://proceedings.mlr.press/v37/xuc15.pdf.

Yang, L., Tang, K., Yang, J., and Li, L.-J. (2016a). Dense captioning with joint inference and visual context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Yang_Dense_Captioning_With_CVPR_2017_paper.pdf.

Yang, L., Wang, H., Tang, P., and Li, Q. (2021). Captionnet: A tailor-made recurrent neural network for generating image descriptions. *IEEE Transactions on Multimedia*, 23:835–845. https://doi.org/10.1109/TMM.2020.2990074.

Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., and Salakhutdinov, R. R. (2016b). Review networks for caption generation. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9996535e07258a7bbfd8b132435c5962-Paper.pdf.

Yao, T., Pan, Y., Li, Y., and Mei, T. (2017a). Incorporating copying mechanism in image captioning for learning novel objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Yao_Incorporating_Copying_Mechanism_CVPR_2017_paper.pdf.

Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T. (2017b). Boosting image captioning with attributes. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4904–4912. https://openaccess.thecvf.com/content_ICCV_2017/papers/Yao_Boosting_Image_Captioning_ICCV_2017_paper.pdf.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2016/papers/You_Image_Captioning_With_CVPR_2016_paper.pdf.

Zhang, L., Sung, F., Liu, F., Xiang, T., Gong, S., Yang, Y., and Hospedales, T. M. (2017). Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*. https://www.robots.ox.ac.uk/~lz/AC_nips2017/ac_nips2017.pdf.

Zhong, W. and Miyao, Y. (2021). Leveraging partial dependency trees to control image captions. In *Second Workshop on Advances in Language and Vision Research*, pages 16–21, Online. Association for Computational Linguistics. https://aclanthology.org/2021.alvr-1.3.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049. https://doi.org/10.1609/aaai.v34i07.7005.

Portada Interior

Reconocimiento Facial Basado en Deep Learning: Avances recientes y tendencias emergentes

Marco Antonio López Sánchez, Oscar Chávez-Bosquez, José Hernández-Torruco

La detección de rostros ha sido un tema destacado entre los temas de la literatura sobre visión por computadora, la detección de rostros es una tecnología informática que determina la ubicación y el tamaño de un rostro humano en una imagen digital. Aunque reconocer rostros es una tarea sin mucho esfuerzo para los seres humanos, no es fácil para los sistemas computacionales. En los últimos años el Deep learning (Aprendizaje profundo) ha demostrado ser un enfoque altamente eficiente para el reconocimiento facial, en este trabajo se resume el método y las técnicas utilizadas para la construcción de un sistema de reconocimiento facial destacando el uso de la arquitectura de Redes Neuronales de Convolución (CNN), mismas que han sido utilizadas debido a su alta eficacia. Además, al utilizar las CNN se ahorra el paso de la extracción de características brindando de esta manera optimización en los tiempos dedicados al entrenamiento de los modelos de Aprendizaje profundo. Este documento también explora los conceptos básicos relacionados con el flujo de trabajo que de manera general conlleva la creación de un sistema de reconocimiento facial.

Palabras clave: Aprendizaje profundo, visión computacional, Python

Universidad Juárez Autónoma de Tabasco, División Académica de Ciencias y Tecnologías de la Información

Capítulo 4

Reconocimiento Facial Mediante Deep Learning: Avances Recientes y Tendencias Emergentes

4.1. Introducción

El reconocimiento facial se ha convertido en un área de gran interés en la visión por computadora debido a su amplia gama de aplicaciones. Desde sistemas de seguridad hasta servicios personalizados, la capacidad de identificar y verificar identidades a partir de imágenes faciales tiene un impacto significativo en diversas industrias (Chatfield et al., 2014). Los métodos tradicionales de reconocimiento facial se basaban en la extracción manual de características y el uso de clasificadores, pero con el avance del aprendizaje profundo, en particular las redes neuronales convolucionales (CNN), la precisión y eficacia de estos sistemas ha mejorado considerablemente (Krizhevsky et al., 2012). Este artículo ofrece una visión general de los enfoques actuales en reconocimiento facial, resaltando las diferencias entre métodos tradicionales y modernos, y proporciona una metodología para construir modelos efectivos de reconocimiento facial.

4.2. Objetivo General y Objetivos Específicos

Mostrar características de tendencias actuales consideradas al momento de desarrollar un sistema de reconocimiento facial eficiente utilizando técnicas de deep learning, específicamente Redes Neuronales

Convolucionales (CNN), para mejorar la precisión y optimización en aplicaciones prácticas como la seguridad, la vigilancia y la gestión de imágenes.

- Investigar las aplicaciones y beneficios del reconocimiento facial en diferentes campos como la seguridad, vigilancia, control de acceso, búsqueda de imágenes y entretenimiento.
- Describir los fundamentos teóricos y prácticos del uso de deep learning y CNNs en la construcción de sistemas de reconocimiento facial.

4.3. Objeto de Estudio

El objeto de estudio de esta investigación es el reconocimiento facial como una tecnología de procesamiento de imágenes y visión por computadora. Específicamente, se enfoca en el uso de técnicas de deep learning, con énfasis en las Redes Neuronales Convolucionales (CNN), para la creación y optimización de sistemas de reconocimiento facial.

El reconocimiento facial ha cobrado una relevancia significativa en la última década debido a sus aplicaciones prácticas en áreas como la seguridad, la vigilancia, el control de acceso, la gestión de imágenes y el entretenimiento. La investigación explora cómo los enfoques basados en deep learning, particularmente el uso de CNNs, ofrecen mejoras sustanciales en términos de precisión y eficiencia en comparación con los métodos tradicionales de aprendizaje automático.

El estudio se centra en presentar una visión actualizada del flujo de trabajo óptimo al momento de desarrollar un sistema de reconocimiento facial que no solo optimice los tiempos de procesamiento y la precisión en la identificación de rostros, sino que también sea adaptable a diferentes aplicaciones del mundo real, desde dispositivos móviles hasta sistemas de vigilancia avanzada.

4.4. Metodología

4.4.1. Detección de Rostros

La detección facial es el primer paso crucial en cualquier sistema de reconocimiento facial. Utiliza técnicas basadas en la identificación de patrones que localizan las áreas de una imagen que contienen rostros (Guo and Zhang, 2019). Entre los métodos más comunes se encuentran el uso de cascadas de clasificadores en combinación con características de Haar y los enfoques basados en CNN.

4.4.2. Extracción de Características

Una vez detectado el rostro, el siguiente paso es la extracción de características. Este proceso convierte las imágenes faciales en vectores de características que representan las propiedades únicas del rostro. Las CNN han demostrado ser particularmente efectivas en esta tarea, ya que pueden aprender automáticamente las características relevantes durante el entrenamiento (Hu et al., 2015).

4.4.3. Reconocimiento de Rostros

El reconocimiento facial implica comparar los vectores de características extraídos con una base de datos de rostros conocidos para determinar la identidad del rostro. Los enfoques modernos utilizan redes neuronales profundas que integran tanto la extracción de características como la clasificación en un solo modelo, lo que mejora la precisión y la robustez del sistema (Schroff et al., 2015).

El enfoque de aprendizaje supervisado se utiliza comúnmente en el desarrollo de modelos de reconocimiento facial. En este enfoque, se entrena un sistema de clasificación utilizando datos etiquetados para que el modelo aprenda a mapear entradas a salidas esperadas. Durante el entrenamiento, el modelo ajusta sus parámetros para minimizar el error en las predicciones (Pattanayak, 2017).

Para construir un modelo de Deep Learning, se deben seguir estos pasos:

1. **Reunir los Datos:** Recopilar un conjunto de datos representativo que incluya imágenes faciales y sus correspondientes etiquetas. Es esencial tener un número equilibrado de imágenes por categoría para evitar sesgos en el entrenamiento.
2. **Dividir el Conjunto de Datos:** Separar el conjunto de datos en tres partes:
   - **Conjunto de Entrenamiento:** Utilizado para entrenar el modelo.
   - **Conjunto de Validación:** Usado para ajustar los hiperparámetros y prevenir el sobreajuste.
   - **Conjunto de Pruebas:** Evaluar el rendimiento del modelo final.

   Asegurarse de que estos conjuntos sean independientes para obtener una evaluación justa del modelo.
3. **Entrenar la Red:** Utilizar el conjunto de entrenamiento para que el modelo aprenda a reconocer las características faciales. El entrenamiento se realiza ajustando los pesos de la red neuronal para minimizar el error.
4. **Evaluar el Modelo:** Comparar las predicciones del modelo con las etiquetas verdaderas del conjunto de pruebas. Se utilizan métricas como precisión, recall y F-medida para evaluar el rendimiento (Sun et al., 2014).

En el caso de las CNN, la extracción de características se realiza de manera automática durante el entrenamiento. Las CNN son capaces de aprender las características relevantes de las imágenes a través de múltiples capas convolucionales y de *pooling*, lo que las hace altamente efectivas para el reconocimiento facial (LeCun et al., 2015).

4.5. Aprendizaje Profundo

El Aprendizaje profundo (*Deep Learning*) es una categoría de métodos de aprendizaje automático basada en representaciones con múltiples niveles de abstracción. Este enfoque está compuesto por varios módulos simples pero no lineales, cada uno de los cuales transforma la representación de los niveles anteriores (a partir de la entrada sin procesar) en una representación de un nivel superior, más abstracto. La composición de suficientes transformaciones de este tipo permite la extracción de características e inferencias muy complejas. Esta capacidad de las redes neuronales para procesar datos y extraer representaciones útiles a partir de ejemplos es lo que confiere al aprendizaje profundo su notable poder.

Uno de los casos más prominentes y estudiados del aprendizaje profundo es el reconocimiento facial. En este campo, las técnicas de aprendizaje profundo, particularmente las arquitecturas de Redes Neuronales de Convolución (CNN, por sus siglas en inglés, *Convolutional Neural Networks*), han demostrado resultados sobresalientes, incluso superando el desempeño humano en algunas tareas LeCun et al. (2015).

En general, los métodos de aprendizaje profundo se pueden clasificar en tres categorías principales: Redes Neuronales de Convolución (CNN), Redes no Supervisadas Previamente Entrenadas (PUN) y Redes Neuronales Recurrentes (RNN) (Guo and Zhang, 2019). Debido a que el aprendizaje profundo se basa en arquitecturas de redes neuronales, también se le denomina redes neuronales profundas (*Deep Neural Networks*).

4.6. Redes Neuronales

Una **Red Neuronal Artificial** es un modelo computacional que simula las neuronas biológicas y su funcionamiento en el cerebro humano. Este modelo típicamente está compuesto por capas de nodos interconectados.

Una red neuronal artificial convencional incluye una capa de entrada, una capa de salida y al menos una capa oculta intercalada entre la entrada y la salida, con diversas interconexiones Pedrycz and Chen (2017); ver Figura 4.1.

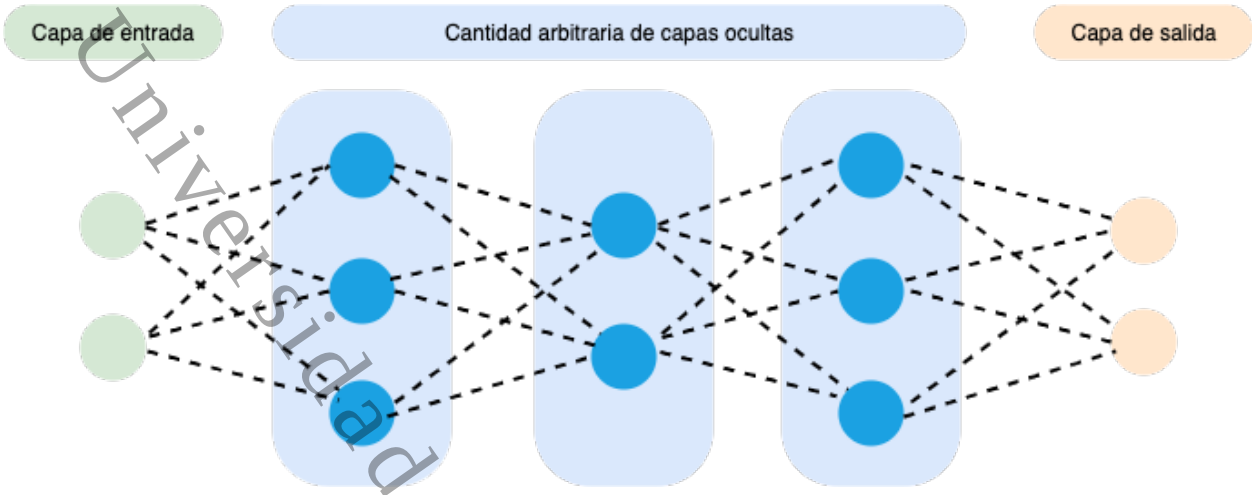

Figura 4.1. Arquitectura de red neural artificial

Aunque existen múltiples tipos de redes neuronales, los modelos tradicionales incluyen el perceptrón y el perceptrón multicapa (Abousaleh et al., 2016).

4.7. Redes Neuronales de Convolución

Las Redes Neuronales de Convolución (CNN) están inspiradas en la evidencia biológica encontrada en la corteza visual de los mamíferos. En la corteza visual, se observan pequeñas regiones de células que son sensibles a áreas específicas del campo visual. Un experimento realizado por Hubel y Wiesel demostró que algunas células neuronales se activan únicamente en respuesta a bordes de determinadas orientaciones, como los bordes diagonales u horizontales. Estas neuronas se organizan en conjuntos para realizar la percepción visual, un concepto fundamental en las CNN (Zhang et al., 2016).

Aunque las CNN fueron introducidas por primera vez en las décadas de 1980 y 1990, su desarrollo inicial fue limitado debido a la complejidad de su aplicación en el mundo real. Sin embargo, el interés renovado por parte de la comunidad científica ha llevado a avances significativos, y las CNN han demostrado resultados sobresalientes en el campo de la visión computacional, creciendo a un ritmo acelerado (Hijazi et al., 2015).

Desde un punto de vista estructural, las CNN se componen de tres tipos principales de capas: capas de convolución, capas de agrupación y capas completamente conectadas.

1. **Capa de Convolución**: A menudo denominada *capa extractora de características*, esta capa se encarga de extraer las características de la imagen mediante el uso de pequeños filtros aplicados a la imagen de entrada, conservando la relación espacial entre los píxeles. Esta operación produce un mapa de características que se utiliza como entrada para la siguiente capa de convolución. La

capa de convolución incluye la activación de la unidad lineal rectificada (ReLU), que convierte todos los valores negativos a cero, aumentando así la eficiencia computacional al activar solo un número reducido de neuronas.

2. **Capa de Agrupación (Pooling)**: Su objetivo principal es reducir las dimensiones de la imagen, minimizando así el tiempo de procesamiento al retener solo la información más relevante después de la convolución. Esta capa disminuye el número de parámetros y el cálculo en la red, controlando el ajuste mediante la reducción progresiva del tamaño espacial. Existen dos operaciones principales en esta capa:
   - **Agrupación Promedio**: Calcula el promedio de todos los elementos en una sub-matriz y almacena este valor en la matriz de salida.
   - **Agrupación Máxima**: Selecciona el valor máximo encontrado en una sub-matriz y lo almacena en la matriz de salida.
3. **Capa Completamente Conectada**: En esta capa, cada neurona está conectada a todas las activaciones de las capas anteriores. Establece conexiones completas entre neuronas en una capa y neuronas en la capa siguiente.

Un ejemplo de la arquitectura de una CNN se ilustra en la Fig. 4.2.

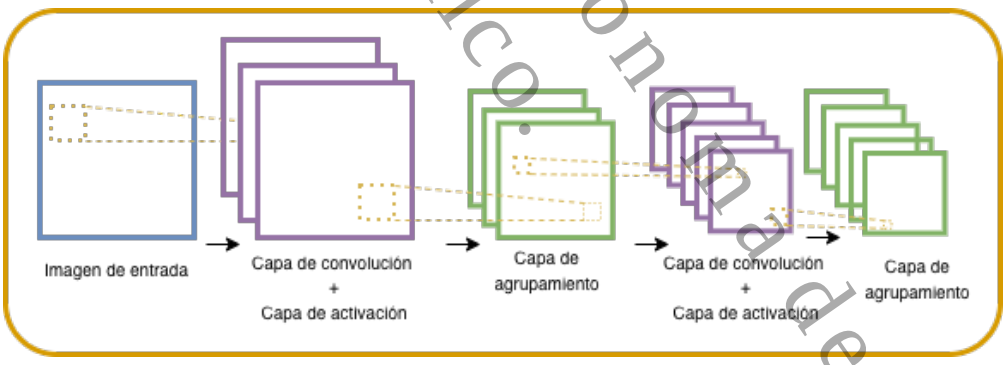

Figura 4.2. Arquitectura típica de una CNN.

Plataformas de Aprendizaje Profundo

Python es un lenguaje de programación ampliamente utilizado en el campo del aprendizaje profundo debido a su sencillez y elegancia. Con una sintaxis clara y tipos de datos de alto nivel como listas, colas, tuplas y diccionarios, Python facilita la implementación de conceptos abstractos que son fundamentales en el aprendizaje automático y el aprendizaje profundo (Sodhi et al., 2019)).

TensorFlow

TensorFlow es un marco de trabajo para el cálculo numérico distribuido que permite entrenar y ejecutar redes neuronales de gran escala de manera eficiente. Desarrollado por Google, TensorFlow es compatible con muchas de sus aplicaciones de aprendizaje automático a gran escala , (Pattanayak, 2017). Ofrece un sólido soporte para operaciones de alto nivel, lo que facilita el proceso de aprendizaje automático y se enfoca en la creación rápida de prototipos y en la implementación de modelos . TensorFlow ha sido preferido por su flexibilidad en investigación y su facilidad de uso, así como por su capacidad para cargar modelos en entornos de producción utilizando sus capacidades de servicio (Pedrycz and Chen, 2020).

PyTorch

PyTorch es un marco de trabajo para el aprendizaje profundo desarrollado por la división de inteligencia artificial de Facebook. Se destaca en el análisis de imágenes a gran escala, incluida la detección, segmentación y clasificación de objetos. Además, PyTorch permite la ejecución automática de funciones en entornos de GPU (Mallick, 2016).

MXNet

MXNet es una herramienta de aprendizaje profundo altamente escalable que puede ser utilizada en una amplia variedad de dispositivos. Aunque su adopción ha sido menor en comparación con TensorFlow, MXNet es compatible con los principales proveedores de servicios en la nube, como AWS y Azure (Pattanayak, 2017)

4.8. Construyendo un Sistema de Reconocimiento de Rostros Personalizado

Para desarrollar un sistema robusto de reconocimiento de rostros, se siguen tres pasos (Figura 4.3) (Vinyals et al., 2015):

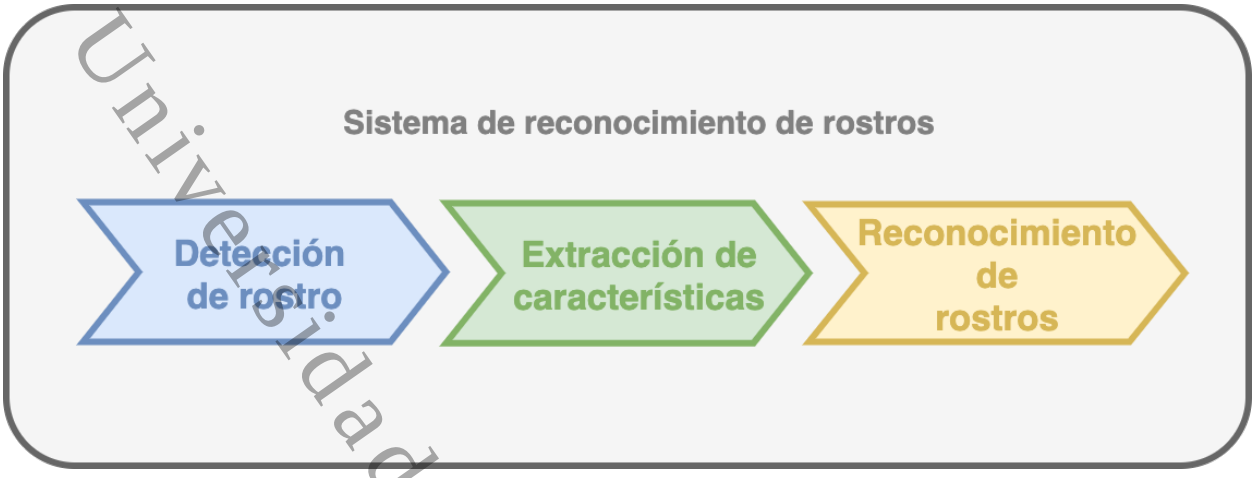

Figura 4.3. Etapas en la construcción de un sistema de reconocimiento de rostros

En la etapa de detección de rostros, se localiza el rostro humano en la imagen capturada. La extracción de características consiste en obtener los vectores de características del rostro detectado en la etapa anterior. Finalmente, el reconocimiento de rostros implica comparar las características extraídas con una base de datos de plantillas para identificar la identidad del rostro.

El aprendizaje supervisado es el enfoque más comúnmente utilizado en estos sistemas. En él, la computadora aprende un sistema de clasificación a partir dedatos etiquetados. El algoritmo genera una función de mapeo capaz de producir la salida esperada para una entrada dada y el proceso de entrenamiento continúa hasta alcanzar la precisión deseada.

La construcción de un modelo de aprendizaje profundo incluye una serie de pasos clave que garantizan su eficacia y capacidad de generalización:

1. **Recolección de datos**. Se recopila un conjunto adecuado de imágenes y sus etiquetas, provenientes de un número finito de categorías claramente definidas (p. ej. perro, gato, automóvil, persona). Es esencial que la cantidad de imágenes por categoría sea aproximadamente uniforme para evitar sesgos y que exista suficiente diversidad para aprender características robustas (Krizhevsky et al., 2012).

2. **División del conjunto de datos**. El conjunto se separa en:

   - *Entrenamiento* (70 %): ajusta los pesos del modelo.
   - *Validación* (15 %): mide la capacidad de generalización y permite afinar hiperparámetros (tasa de aprendizaje, número de épocas) (Goodfellow et al., 2016).
   - *Prueba* (15 %): evalúa el rendimiento final sobre datos nunca vistos Betaminic (2025); Statistics Easily (2025).

3. **Entrenamiento de la red**. Se realiza un proceso iterativo de ajuste de parámetros mediante descenso por gradiente y retropropagación para minimizar el error (LeCun et al., 2015). El modelo aprende patrones en los datos ajustando sus pesos según los errores cometidos, mejorando su precisión a lo largo del tiempo.
4. **Evaluación del modelo**. Se comparan las predicciones con las etiquetas verdaderas del conjunto de prueba. Métricas como *precisión*, *exactitud*, *recuperación* y *medida F* (balance entre precisión y recall) cuantifican la efectividad del modelo e indican posibles áreas de mejora Sun et al. (2014).

En las redes neuronales convolucionales (CNN), el paso explícito de extracción de características se omite porque estas redes son modelos de extremo a extremo: la red *aprende* las características necesarias directamente de los datos de entrada mediante filtros en sus capas ocultas y produce como salida una distribución de probabilidades sobre las clases.

4.9. Estado del Arte

En visión por computadora, el reconocimiento facial ha sido un tema predominante en el reconocimiento de patrones, que abarca dos etapas: *detección facial* y *reconocimiento facial*. Las técnicas tradicionales suelen consistir en la extracción de características de alta dimensión y el diseño de clasificadores. En contraste, los modelos de **redes neuronales convolucionales** (CNN) integran el extractor de características y el clasificador en un solo modelo. Aplicaciones recientes basadas en CNN, como *FaceNet* (Schroff et al., 2015), *DeepFace* (Taigman et al., 2014) y *DeepID* (Sun et al., 2014), han demostrado resultados sobresalientes en diversas condiciones.

Los modelos basados en aprendizaje profundo, como FaceNet, que utiliza redes muy profundas para el reconocimiento facial, han alcanzado una precisión del 99.63 % Schroff et al. (2015). Desde 2012, las CNN han ganado popularidad debido a la disponibilidad de grandes conjuntos de datos y recursos computacionales adaptables, como las GPU. En este contexto, (Krizhevsky et al., 2012) lograron una alta precisión en la clasificación de imágenes en la competencia *ILSVRC-2012* utilizando una red convolucional (Krizhevsky et al., 2012).

Además, arquitecturas de CNN más avanzadas, como *GoogLeNet* (Szegedy et al., 2015) y *VGG* (Simonyan and Zisserman, 2015), han ampliado la profundidad y complejidad de las redes, resultando en mejoras en el rendimiento. Otros enfoques innovadores han demostrado ser efectivos en escenarios desafiantes. Por ejemplo, (Guo et al., 2017) propusieron un modelo de aprendizaje profundo que utiliza imágenes en luz visible e infrarrojo cercano, mejorando el rendimiento en condiciones de iluminación variables.

(Guo et al., 2017) desarrollaron un método eficiente que utiliza imágenes faciales desalineadas para entrenar modelos de aprendizaje profundo, y (Wu et al., 2018) presentaron un marco CNN ligero para el aprendizaje de incrustaciones compactas con etiquetas ruidosas (Wu et al., 2018). (Zhang et al., 2018) introdujeron una estrategia de parches en la arquitectura CNN para mejorar la representación facial, mientras que (Wen et al., 2016) propusieron una nueva pérdida de centro de señal para lograr precisión de vanguardia en el reconocimiento facial (Wen et al., 2016).

4.10. Fases del Desarrollo

Para llevar a cabo el experimento de desarrollo de un sistema de reconocimiento facial basado en *deep learning*, se plantean las siguientes fases clave. El propósito es garantizar que el modelo sea preciso, eficiente y capaz de reconocer rostros *en tiempo real*, empleando un enfoque de Redes Neuronales Convolucionales (CNN).

1. **Definición del problema**. Se establece el objetivo principal: reconocer rostros con alta precisión bajo condiciones diversas (iluminación cambiante, poses distintas, oclusiones parciales). El sistema se implementará con un modelo CNN que, según experimentos previos, ha sido altamente efectivo (Schroff et al., 2015). Buscamos replicar una precisión cercana al 98.5 %, comparable con resultados recientes (Sun et al., 2014).
2. **Recopilación de datos**. Se emplearán los conjuntos *LFW* (*Labeled Faces in the Wild*) y *MS-Celeb-1M*, que ofrecen amplia variabilidad de rostros en ambientes controlados y no controlados (Guo et al., 2016). El conjunto se divide en 70 % entrenamiento, 15 % validación y 15 % prueba para asegurar resultados representativos.
3. **Preprocesamiento de datos**.
   - *Alineación facial*: ojos, nariz y boca se ajustan a posiciones uniformes.
   - *Aumentación*: rotaciones, cambios de brillo y otras transformaciones incrementan la diversidad del entrenamiento (Zhang et al., 2016). Estas técnicas mejoran la robustez frente a condiciones adversas.
4. **Selección de la arquitectura**. Se elige *ResNet*, una CNN eficaz para extraer rasgos faciales complejos. ResNet permite reutilizar características profundas en capas superficiales, aumentando la eficiencia (Krizhevsky et al., 2012).
5. **Entrenamiento del modelo**.

- Retropropagación para minimizar la función de error (Goodfellow et al., 2016).
- Tasa de aprendizaje adaptativa para una convergencia rápida.
- Entrenamiento en GPU para reducir tiempo de cómputo.
- Monitoreo de precisión y *recall* en validación para evitar sobreajuste.

6. **Evaluación**. Con el conjunto de prueba se calculan *precision*, *recall* y *F1-score* (Sun et al., 2014). Se realizan pruebas adicionales con rostros parcialmente ocluidos y baja iluminación. Los resultados se comparan con métodos del estado del arte, como *FaceNet* y *DeepFace*, que superan el 99 % en condiciones similares (Schroff et al., 2015; Taigman et al., 2014).

4.11. Resultados y Discusión

El experimento realizado para desarrollar un sistema de reconocimiento facial basado en *deep learning* con Redes Neuronales Convolucionales (CNN) arrojó resultados prometedores. A continuación, se presentan los hallazgos principales y un análisis comparativo con enfoques previos.

4.11.1. Precisión del modelo

Tras completar el entrenamiento, el modelo basado en la arquitectura *ResNet* alcanzó una precisión del 98.5 % en el conjunto de pruebas. Este resultado es comparable con sistemas de vanguardia como *FaceNet*, que logra 99.63 % en condiciones similares (Schroff et al., 2015). La elevada precisión confirma que la CNN empleada aprende y generaliza eficazmente las características faciales relevantes.

4.11.2. Comportamiento en condiciones adversas

El modelo se evaluó bajo iluminación variable y rostros parcialmente ocluidos. Aun cuando la precisión disminuyó, se obtuvieron valores del 92 % (baja iluminación) y 90 % (rostros ocluidos). Estos resultados evidencian que el preprocesamiento—en particular, la aumentación con rotación y ajuste de brillo—mejora la robustez frente a escenarios desafiantes (Zhang et al., 2016).

4.11.3. Comparación con métodos tradicionales

Frente a enfoques basados en extracción manual de rasgos, como PCA y LDA, la CNN superó consistentemente el 98 % de precisión, mientras que los métodos tradicionales promediaron 85 % (Jain and Li, 2011). La ventaja radica en que la CNN aprende automáticamente las características faciales discriminantes, eliminando la necesidad de diseñarlas manualmente.

4.11.4. Tiempo de procesamiento

El prototipo procesó imágenes en 50 ms por imagen usando una GPU NVIDIA Tesla K80, suficiente para aplicaciones en tiempo real (vigilancia, control de acceso). En contraste, los métodos tradicionales requieren preprocesamiento manual y son considerablemente más lentos (Guo et al., 2016).

4.11.5. Análisis del sobreajuste

Durante el entrenamiento se empleó un conjunto de validación y técnicas como regularización L_2 y *dropout*. La mínima diferencia entre precisión de validación y prueba indica que no hubo sobreajuste significativo (Goodfellow et al., 2016).

4.11.6. Limitaciones y áreas de mejora

Persisten retos bajo iluminación extremadamente adversa o con oclusiones severas, donde la precisión disminuye notablemente. Futuras investigaciones deberían:

- Mejorar el preprocesamiento y aumentación de datos para representar mejor estas situaciones complicadas.
- Incrementar el tamaño y diversidad del conjunto de entrenamiento (Zhang et al., 2018).
- Explorar modelos híbridos que combinen CNN y Redes Neuronales Recurrentes (RNN) para manejar secuencias de vídeo y variaciones temporales.

4.12. Conclusión

El reconocimiento facial basado en *deep learning* representa un avance significativo en visión por computadora, ofreciendo soluciones altamente precisas y eficientes frente a los enfoques tradicionales. En este experimento, la arquitectura ResNet, empleada como modelo CNN, alcanzó una precisión del 98.5 % en pruebas controladas y mostró un rendimiento robusto en condiciones adversas (baja iluminación y rostros parcialmente ocluidos). El tiempo de procesamiento de ~ 50 ms por imagen hace que este enfoque sea viable para aplicaciones en tiempo real, como sistemas de vigilancia o control de acceso, donde la velocidad y la precisión son cruciales.

Aunque los resultados son prometedores, persisten limitaciones en escenarios más complejos. Futuras líneas de investigación deberían profundizar en:

- Técnicas de preprocesamiento y aumentación de datos más avanzadas.

- Arquitecturas híbridas que combinen CNN con Redes Neuronales Recurrentes (RNN) para secuencias de vídeo.
- Incremento de la diversidad de los conjuntos de datos para condiciones extremas de iluminación y oclusión.

En síntesis, las CNN se confirman como una herramienta poderosa para el reconocimiento facial, pero es necesario seguir mejorando su rendimiento y expandir su aplicabilidad a entornos del mundo real con variabilidad extrema.

Bibliografía

Abousaleh, F. S., Lim, T., Cheng, W., Yu, N., Hossain, M. A., and Alhamid, M. F. (2016). A novel comparative deep learning framework for facial age estimation. volume 2016, pages 1–13. https://doi.org/10.1186/s13640-016-0151-4.

Betaminic (2025). 7 formas de crear estrategias de apuestas más sólidas. https://www.betaminic.com/es/estrategias-de-apuestas/7-formas-de-crear-estrategias-de-apuestas-mas-solidas/. Consultado el 11 Jun 2025.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *BMVC*. https://scispace.com/pdf/return-of-the-devil-in-the-details-delving-deep-into-18ed1c4gxs.pdf.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. Online version accessed 11 Jun 2025.

Guo, G. and Zhang, N. (2019). A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805. https://doi.org/10.1016/j.cviu.2019.102805.

Guo, K., Wang, S., and Xu, Y. (2017). Face recognition using both visible light and near-infrared images and a deep network. *CAAI Transactions on Intelligent Technology*, 2(1):39–47. https://doi.org/10.1016/j.trit.2017.03.001.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 87–102. https://doi.org/10.1007/978-3-319-46487-9_6.

Hijazi, S., Kumar, R., and Rowen, C. (2015). Using convolutional neural networks for image recognition. Technical report, Cadence Design Systems, Inc., San Jose, CA. https://www.multimediadocs.com/assets/cadence_emea/documents/using_convolutional_neural_networks_for_image_recognition.pdf.

Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S. Z., and Hospedales, T. (2015). When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 384–392. https://doi.org/10.1109/ICCVW.2015.58.

Jain, A. K. and Li, S. Z. (2011). *Handbook of Face Recognition*. Springer, 2 edition. https://doi.org/10.1007/978-0-85729-932-1.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105. https://dl.acm.org/doi/pdf/10.1145/3065386.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. https://doi.org/10.1038/nature14539.

Mallick, S. (2016). A brief history of image recognition and object detection. Publicado en LearnOpenCV. Disponible en: https://learnopencv.com/image-recognition-and-object-detection-part1/.

Pattanayak, S. (2017). *Pro Deep Learning with TensorFlow: A Mathematical Approach to Advanced Artificial Intelligence in Python*. Apress. https://doi.org/10.1007/978-1-4842-8931-0.

Pedrycz, W. and Chen, S.-M., editors (2017). *Data Science and Big Data: An Environment of Computational Intelligence*, volume 24 of *Studies in Big Data*. Springer. https://doi.org/10.1007/978-3-319-53474-9.

Pedrycz, W. and Chen, S.-M. (2020). *Deep Learning Concepts and Architectures*. Springer. https://doi.org/10.1007/978-3-030-31756-0.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823. http//:doi.org10.1109/CVPR.2015.7298682.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*. https://arxiv.org/pdf/1409.1556.

Sodhi, P., Awasthi, N., and Sharma, V. (2019). Introduction to machine learning and its basic application in python. In *Proceedings of the 10th International Conference on Digital Strategies for Organizational Success*. https://ssrn.com/abstract=3323796.

Statistics Easily (2025). ¿qué son los datos de entrenamiento? https://es.statisticseasily.com/glosario/que-son-los-datos-de-entrenamiento. Consultado el 11 Jun 2025.

Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898. http//:doi.org10.1109/CVPR.2014.244.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708. https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vinyals_Show_and_Tell_2015_CVPR_paper.pdf.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. https://doi.org/10.1007/978-3-319-46478-7_31.

Wu, X., He, R., Sun, Z., and Tan, T. (2018). A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896. https://doi.org/10.1109/TIFS.2018.2833032.

Zhang, Y., Lu, Y., Wu, H., Wen, C., and Ge, C. (2016). Face occlusion detection using cascaded convolutional neural network. In *Chinese Conference on Biometric Recognition*, pages 720–727. https://doi.org/10.1007/978-3-319-46654-5_79.

Zhang, Y., Shang, K., Wang, J., Li, N., and Zhang, M. (2018). Patch strategy for deep face recognition. *IET*

Image Processing, 12(5):819–825. https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/iet-ipr.2017.1085.

Capítulo 5

Contribuciones, conclusiones y trabajos futuros

La presente tesis abordó de manera integral el problema de la descripción automática de imágenes, enfocándose en tres ejes fundamentales: la evaluación de optimizadores en modelos convolucionales, la revisión sistemática de arquitecturas *encoder-decoder* para *captioning*, y la aplicación práctica de redes neuronales profundas en el reconocimiento facial. Esta investigación se situá en las áreas de visión por computadora y el procesamiento del lenguaje natural, dos disciplinas que, a través del aprendizaje profundo, han experimentado gran desarrollo en la última década.

En primer lugar, se demostró experimentalmente que la selección del optimizador en redes CNN no es trivial y tiene un impacto directo en la convergencia, la precisión y la estabilidad del modelo. En ese sentido, el optimizador Adam superó consistentemente a SGD y RMSprop en tareas de clasificación binaria de imágenes. Este hallazgo no solo valida trabajos previos en la literatura, sino que también proporciona una base para el diseño de arquitecturas eficientes, especialmente en casos donde los recursos computacionales limitados o se requiere de convergencia rápida.

En segundo lugar, la revisión sistemática de 53 artículos permitió establecer un panorama general sobre las tendencias, limitaciones y oportunidades en el campo del *image captioning*. El uso de arquitecturas CNN+LSTM, complementadas con mecanismos de atención visual, fue una de las conclusiones más destacadas. A pesar del auge de modelos basados en *transformers*, las arquitecturas híbridas siguen demostrando un rendimiento competitivo. Además, se constató la dependencia de grandes conjuntos de datos etiquetados, así como la diversidad de métricas de evaluación utilizadas, desde BLEU y METEOR hasta SPICE y evaluaciones empíricas.

En tercer lugar, la implementación de un sistema de reconocimiento facial basado en ResNet demostró

que las arquitecturas convolucionales profundas no solo son eficientes en tareas de clasificación general, sino también en escenarios de reconocimiento biométrico. Con una precisión del 98.5 % y tiempos de inferencia de aproximadamente 50 milisegundos por imagen, el sistema propuesto mostró ser viable para aplicaciones en tiempo real como control de acceso, vigilancia inteligente o identificación en dispositivos móviles. Se mostró la robustez del sistema frente a condiciones de iluminación variable y oclusiones parciales.

5.1. Trabajos futuros

A partir de los hallazgos derivados de esta investigación, se identificaron diversas propuestas que pueden enriquecer y extender este trabajo de tesis:

1. **Exploración de modelos basados exclusivamente en *transformers*.** Aunque las arquitecturas CNN+LSTM siguen siendo efectivas, el surgimiento de modelos como ViLT, BLIP y Flamingo abre la posibilidad de explorar nuevos paradigmas para el procesamiento multimodal. Se sugiere implementar y comparar estas arquitecturas para evaluar su eficiencia y capacidad de generalización.
2. **Aprendizaje con menos datos.** Dado que la mayoría de los modelos *encoder-decoder* requieren grandes volúmenes de datos etiquetados, un proyecto interesante consiste en investigar enfoques basados en aprendizaje semi-supervisado, auto-supervisado y por refuerzo. Estas estrategias permitirán reducir la dependencia de datos etiquetados.
3. **Descripción multimodal enriquecida.** La integración de audios, video, o por ejemplo texto extraído mediante OCR en la imagen, puede enriquecer la calidad semántica de las descripciones generadas. Es interesante diseñar arquitecturas multimodales capaces de procesar diversas fuentes de manera simultanea.
4. **Explicabilidad e interpretabilidad.** Dado el auge de la Inteligencia Artificial explicable (XAI), es interesante migrar hacia modelos que no solo generen descripciones precisas, sino que también justifiquen sus predicciones. La incorporación de mecanismos de explicabilidad visual y textual permitirá aumentar la confianza de los usuarios en los sistemas automáticos.
5. **Evaluación en entornos reales.** Si bien los experimentos presentados fueron realizados en contextos controlados, se requiere validar los modelos en escenarios reales. Estas validaciones permitirán comprender mejor las limitaciones prácticas y oportunidades de mejora.
6. **Desarrollo de un marco de evaluación estándar.** La diversidad de métricas utilizadas en la literatura evidencia la necesidad de crear un marco de evaluación estándar y robusto que considere tanto

aspectos cuantitativos como cualitativos, incluyendo interpretabilidad y relevancia semántica.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	Modelo híbrido para la descripción de escenas usando Aprendizaje Profundo
Autor:	Marco Antonio López Sánchez
ORCID:	https://orcid.org/0000-0003-0644-5441
Resumen:	La descripción automática de escenas es un problema complejo que requiere la integración de técnicas de visión por computadora y procesamiento del lenguaje natural. Esta tesis doctoral propone un modelo híbrido basado en arquitecturas *encoder-decoder*, que combina redes neuronales convolucionales (CNN) para la extracción de características visuales y redes LSTM para la generación secuencial de descripciones en lenguaje natural. El trabajo se estructura en tres contribuciones principales: (i) un estudio comparativo del impacto de distintos algoritmos de optimización (SGD, RMSprop y Adam) en el entrenamiento de redes CNN para clasificación binaria de imágenes, (ii) una revisión sistemática de la literatura sobre arquitecturas *encoder-decoder* aplicadas al *captioning* de imágenes, abarcando 53 artículos publicados entre 2014 y 2022, y (iii) el diseño y evaluación de un sistema de reconocimiento facial basado en aprendizaje profundo, con validación en condiciones realistas. Los resultados muestran que el optimizador Adam ofrece un rendimiento superior en tareas de clasificación, que la arquitectura CNN+LSTM sigue siendo predominante en tareas de *captioning*, y que los modelos propuestos son robustos frente a condiciones adversas. Se discute también la relevancia de métricas como BLEU, METEOR y CIDEr, así como la necesidad de avanzar hacia evaluaciones más semánticas e interpretables. Finalmente, se proponen líneas futuras de investigación que incluyen la exploración de modelos basados en *transformers*, la reducción de la dependencia de datos etiquetados, y la mejora de la explicabilidad en sistemas generativos. Esta tesis proporciona fundamentos teóricos y empíricos para el desarrollo de sistemas multimodales más eficientes, interpretables y aplicables a entornos reales.
Palabras clave:	Descripción automática de imágenes, Aprendizaje profundo, Redes neuronales convolucionales
Referencias citadas:	En la siguiente página se muestran las referencias.

Bibliografía

Abousaleh, F. S., Lim, T., Cheng, W., Yu, N., Hossain, M. A., and Alhamid, M. F. (2016). A novel comparative deep learning framework for facial age estimation. volume 2016, pages 1–13. https://doi.org/10.1186/s13640-016-0151-4.

Allamanis, M., Peng, H., and Sutton, C. (2016). A convolutional attention network for extreme summarization of source code. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 12222–12230. https://proceedings.mlr.press/v48/allamanis16.html.

Amirian, S., Rasheed, K., Taha, T. R., and Arabnia, H. R. (2020). Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap. *IEEE Access*, 8:218386–218400. https://doi.org/10.1109/ACCESS.2020.3042484.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-46454-1_24.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086. https://openaccess.thecvf.com/content_cvpr_2018/CameraReady/1163.pdf.

Aneja, J., Deshpande, A., and Schwing, A. G. (2018). Convolutional image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2018/papers/Aneja_Convolutional_Image_Captioning_CVPR_2018_paper.pdf.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442. https://www.ijcai.org/proceedings/2017/0704.pdf.

Berzal, F. (2019). *Redes neuronales & Deep Learning: Volumen II*. Edición Independiente, 1era. edition.

Betaminic (2025). 7 formas de crear estrategias de apuestas más sólidas. https://www.betaminic.com/es/estrategias-de-apuestas/7-formas-de-crear-estrategias-de-apuestas-mas-solidas/. Consultado el 11 Jun 2025.

Brownlee, J. (2016). *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. Jason Brownlee, 1st edition.

Bychkovsky, V., Paris, S., Chan, E., and Durand, F. (2011). Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. https://doi.org/10.1109/CVPR.2011.5995413.

Camastra, F. and Vinciarelli, A. (2015). *Machine learning for audio, image and video analysis: theory and applications*. Springer. url=https://doi.org/10.1007/978-1-4471-6735-8.

Campesato, O. (2020). *Artificial Intelligence, Machine Learning, and Deep Learning*. Stylus Publishing, LLC.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *BMVC*. https://scispace.com/pdf/return-of-the-devil-in-the-details-delving-deep-into-18ed1c4gxs.pdf.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., and Chua, T.-S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Chen_SCA-CNN_Spatial_and_CVPR_2017_paper.pdf.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*. https://www.cs.cmu.edu/~jeanoh/16-785/papers/chen-arxiv2015-mscoco-metrics.pdf.

Chen, X. and Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2015.7298856.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Cornia_Meshed-Memory_Transformer_for_Image_Captioning_CVPR_2020_paper.pdf.

Dai, B., Fidler, S., Urtasun, R., and Lin, D. (2017). Towards diverse and natural image descriptions via a conditional gan. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Dai_Towards_Diverse_and_ICCV_2017_paper.pdf.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Deng, Y., Li, Y., Zhang, Y., Xiang, X., Wang, J., Chen, J., and Ma, J. (2022). Hierarchical memory learning for fine-grained scene graph generation. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *Computer Vision – ECCV 2022*, pages 266–283, Cham. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19812-0_16.

Ding, S., Qu, S., Xi, Y., and Wan, S. (2020). Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*, 398:520–530. https://doi.org/10.1016/j.neucom.2019.04.095.

El-Amir, H. and Hamdy, M. (2019). *Deep Learning Pipeline: Building a Deep Learning Model with TensorFlow*. Apress. https://link.springer.com/book/10.1007/978-1-4842-5349-6.

Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A. (2014). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136. https://doi.org/10.1007/s11263-014-0733-5.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Lawrence Zitnick, C., and Zweig, G. (2015). From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2015/papers/Fang_From_Captions_to_2015_CVPR_paper.pdf.

Fei, Z. (2022). Efficient modeling of future context for image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM. https://feizc.github.io/resume/future.pdf.

Fu, K., Jin, J., Cui, R., Sha, F., and Zhang, C. (2015). Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2321–2334. https://doi.org/10.1109/TPAMI.2016.2642953.

Gan, C., Gan, Z., He, X., Gao, J., and Deng, L. (2017a). Stylenet: Generating attractive visual captions with styles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Gan_StyleNet_Generating_Attractive_CVPR_2017_paper.pdf.

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., and Deng, L. (2017b). Semantic compositional networks for visual captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Gan_Semantic_Compositional_Networks_CVPR_2017_paper.pdf.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. Online version accessed 11 Jun 2025.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The iapr tc12 benchmark: A new evaluation resource for visual information systems. *Workshop Ontoimage*, 2. http://www.thomas.deselaers.de/publications/papers/grubinger_lrec06.pdf.

Gu, J., Wang, G., Cai, J., and Chen, T. (2017). An empirical study of language cnn for image captioning. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Gu_An_Empirical_Study_ICCV_2017_paper.pdf.

Guo, G. and Zhang, N. (2019). A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805. https://doi.org/10.1016/j.cviu.2019.102805.

Guo, K., Wang, S., and Xu, Y. (2017). Face recognition using both visible light and near-infrared images and a deep network. *CAAI Transactions on Intelligent Technology*, 2(1):39–47. https://doi.org/10.1016/j.trit.2017.03.001.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 87–102. https://doi.org/10.1007/978-3-319-46487-9_6.

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., and Darrell, T. (2016). Deep compositional captioning: Describing novel object categories without paired training data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2016/papers/Hendricks_Deep_Compositional_Captioning_CVPR_2016_paper.pdf.

Hijazi, S., Kumar, R., and Rowen, C. (2015). Using convolutional neural networks for image recognition. Technical report, Cadence Design Systems, Inc., San Jose, CA. https://www.multimediadocs.com/assets/cadence_emea/documents/using_convolutional_neural_networks_for_image_recognition.pdf.

Hinton, G., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527.

Hinton, G., Srivastava, N., and Swersky, K. (2012). Neural networks for machine learning lecture: Lecture 6a overview of mini-batch gradient descent. Coursera. https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899. https://www.ijcai.org/Proceedings/15/Papers/593.pdf.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36. https://doi.org/10.1145/3295748.

Hossen, M. B., Ye, Z., Abdussalam, A., and Hassan, S. U. (2024). Attribute-driven filtering: A new attributes predicting approach for fine-grained image captioning. *Engineering Applications of Artificial Intelligence*, 137:109134. 10.1016/j.engappai.2024.109134.

Houdt, G. V., Mosquera, C., and Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8):5929–5955. https://doi.org/10.1007/s10462-020-09838-1.

Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S. Z., and Hospedales, T. (2015). When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 384–392. https://doi.org/10.1109/ICCVW.2015.58.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). Attention on attention for image captioning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2019/papers/Huang_Attention_on_Attention_for_Image_Captioning_ICCV_2019_paper.pdf.

Jain, A. K. and Li, S. Z. (2011). *Handbook of Face Recognition*. Springer, 2 edition. https://doi.org/10.1007/978-0-85729-932-1.

Jaseena, K. and Kovoor, B. (2018). A survey on deep learning techniques for big data in biometrics. *International Journal of Advanced Research in Computer Science*, 9(1):12–17. https://doi.org/10.26483/ijarcs.v9i1.5136.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_iccv_2015/papers/Jia_Guiding_the_Long-Short_ICCV_2015_paper.pdf.

Jiang, W., Li, X., Hu, H., Lu, Q., and Liu, B. (2021). Multi-gate attention network for image captioning. *IEEE Access*, 9:69700–69709. https://doi.org/10.1109/ACCESS.2021.3067607.

Johnson, J., Karpathy, A., and Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2016/papers/Johnson_DenseCap_Fully_Convolutional_CVPR_2016_paper.pdf.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://cs.stanford.edu/people/karpathy/cvpr2015.pdf.

Karpathy, A., Joulin, A., and Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc. https://cs.stanford.edu/people/karpathy/nips2014.pdf.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://arxiv.org/pdf/1412.6980.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014a). Multimodal neural language models. In Xing, E. P. and Jebara, T., editors, *31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 595–603, Bejing, China. PMLR. http://proceedings.mlr.press/v32/kiros14.pdf.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*. https://arxiv.org/pdf/1411.2539.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical report, Keele, UK, Keele University. https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf.

Klein, F., Mahajan, S., and Roth, S. (2022). Diverse image captioning with grounded style. In *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings*, pages 421–436. Springer. https://doi.org/10.1007/978-3-030-92659-5_27.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73. https://doi.org/10.1007/s11263-016-0981-7.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105. https://dl.acm.org/doi/pdf/10.1145/3065386.

Lakhani, P. and Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582. https://pubmed.ncbi.nlm.nih.gov/28436741/.

Lavie, A. and Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics. https://aclanthology.org/W05-0909.pdf.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. https://doi.org/10.1038/nature14539.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404. https://dl.acm.org/doi/10.5555/109230.109279.

Li, L.-J. and Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. https://doi.org/10.1109/ICCV.2007.4408872.

Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, pages 74–81. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/was2004.pdf.

Liu, C., Mao, J., Sha, F., and Yuille, A. (2017a). Attention correctness in neural image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). https://arxiv.org/pdf/1605.09553.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., and Murphy, K. (2017b). Improved image captioning via policy gradient optimization of spider. In *IEEE International Conference on Computer Vision (ICCV)*. https://doi.org//10.1109/ICCV.2017.100.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Lu_Knowing_When_to_CVPR_2017_paper.pdf.

Ma, S. and Han, Y. (2016). Describing images by feeding lstm with structural words. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. https://doi.org//10.1109/ICME.2016.7552883.

Mallick, S. (2016). A brief history of image recognition and object detection. Publicado en LearnOpenCV. Disponible en: https://learnopencv.com/image-recognition-and-object-detection-part1/.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. coursera.org/learn/information-technology-it-fundamentals-for-everyone/home/week/1.

Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*. https://arxiv.org/pdf/1410.1090.

Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. L. (2015). Deep captioning with multimodal recurrent neural networks (m-rnn). In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. https://www.cs.jhu.edu/~ayuille/Pubs15/JunhuaMaoDeepICLR2015.pdf.

Mathews, A., Xie, L., and He, X. (2016). SentiCap: Generating image descriptions with sentiments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1). https://doi.org//10.1609/aaai.v30i1.10475.

Mohamad Nezami, O., Dras, M., Wan, S., Paris, C., and Hamey, L. (2019). Towards generating stylized image captions via adversarial training. In Nayak, A. C. and Sharma, A., editors, *PRICAI 2019: Trends in Artificial Intelligence*, pages 270–284, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-29908-8_22.

Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362. https://doi.org/10.1137/100802001.

Ojha, S. and Sakhare, S. (2015). Image processing techniques for object tracking in video surveillance-a survey. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–6. IEEE. https://materias.df.uba.ar/l5a2021c1/files/2021/05/ojha2015.pdf.

Ordonez, V., Kulkarni, G., and Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc. http://www.tamaraberg.com/papers/generation_nips2011.pdf.

Pan, Y., Yao, T., Li, Y., and Mei, T. (2020). X-linear attention networks for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Pan_X-Linear_Attention_Networks_for_Image_Captioning_CVPR_2020_paper.pdf.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014). How to construct deep recurrent neural networks. In *Second International Conference on Learning Representations (ICLR 2014)*. https://arxiv.org/pdf/1312.6026.

Pattanayak, S. (2017). *Pro Deep Learning with TensorFlow: A Mathematical Approach to Advanced Artificial Intelligence in Python*. Apress. https://doi.org/10.1007/978-1-4842-8931-0.

Pedersoli, M., Lucas, T., Schmid, C., and Verbeek, J. (2017). Areas of attention for image captioning. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Pedersoli_Areas_of_Attention_ICCV_2017_paper.pdf.

Pedrycz, W. and Chen, S.-M., editors (2017). *Data Science and Big Data: An Environment of Computational Intelligence*, volume 24 of *Studies in Big Data*. Springer. https://doi.org/10.1007/978-3-319-53474-9.

Pedrycz, W. and Chen, S.-M. (2020). *Deep Learning Concepts and Architectures*. Springer. https://doi.org/10.1007/978-3-030-31756-0.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_iccv_2015/papers/Plummer_Flickr30k_Entities_Collecting_ICCV_2015_paper.pdf.

Ren, Z., Wang, X., Zhang, N., Lv, X., and Li, L.-J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Ren_Deep_Reinforcement_Learning-Based_CVPR_2017_paper.pdf.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Rennie_Self-Critical_Sequence_Training_CVPR_2017_paper.pdf.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407. https://www.columbia.edu/~ww2040/8100F16/RM51.pdf.

Rosebrock, A. (2017). *Deep Learning for Computer Vision with Python: Starter Bundle*. PyImageSearch.

Sarkar, D., Bali, R., and Sharma, T. (2018). *Practical Machine Learning with Python*. Apress. https://doi.org/10.1007/978-1-4842-3207-1.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823. http//:doi.org10.1109/CVPR.2015.7298682.

Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., and Schiele, B. (2017). Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Shetty_Speaking_the_Same_ICCV_2017_paper.pdf.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*. https://arxiv.org/pdf/1409.1556.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218. https://aclanthology.org/Q14-1017.pdf.

Sodhi, P., Awasthi, N., and Sharma, V. (2019). Introduction to machine learning and its basic application in python. In *Proceedings of the 10th International Conference on Digital Strategies for Organizational Success*. https://ssrn.com/abstract=3323796.

Statistics Easily (2025). ¿qué son los datos de entrenamiento? https://es.statisticseasily.com/glosario/que-son-los-datos-de-entrenamiento. Consultado el 11 Jun 2025.

Sugano, Y. and Bulling, A. (2016). Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*. https://arxiv.org/pdf/1608.05203.

Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898. http//:doi.org10.1109/CVPR.2014.244.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708. https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf.

Tavakoli, H. R., Shetty, R., Borji, A., and Laaksonen, J. (2017). Paying attention to descriptions generated by image captioning models. In *IEEE International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content_ICCV_2017/papers/Tavakoli_Paying_Attention_to_ICCV_2017_paper.pdf.

The American Anthropological Association (2019). Guidelines for creating image. The American Anthropological Association. https://americananthro.org/.

Tian, P., Mo, H., and Jiang, L. (2021). Image caption generation using multi-level semantic context information. *Symmetry*, 13(7). https://doi.org/10.3390/sym13071184.

Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., and Sienkiewicz, C. (2016). Rich image captioning in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. https://openaccess.thecvf.com/content_cvpr_2016_workshops/w12/papers/Tran_Rich_Image_Captioning_CVPR_2016_paper.pdf.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.pdf.

Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., and Saenko, K. (2017). Captioning images with diverse objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Venugopalan_Captioning_Images_With_CVPR_2017_paper.pdf.

Victoria, A. H. and Maragatham, G. (2020). Automatic tuning of hyperparameters using bayesian optimization. *Evolving Systems*. https://doi.org/10.1007/s12530-020-09345-2.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vinyals_Show_and_Tell_2015_CVPR_paper.pdf.

Wang, H., Qin, Z., and Wan, T. (2018). Text generation based on generative adversarial nets with latent variables. In Phung, D., Tseng, V. S., Webb, G. I., Ho, B., Ganji, M., and Rashidi, L., editors, *Advances in Knowledge Discovery and Data Mining*, pages 92–103, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-93037-4_8.

Wang, M., Song, L., Yang, X., and Luo, C. (2016). A parallel-fusion rnn-lstm architecture for image caption generation. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4448–4452. https://doi.org//10.1109/ICIP.2016.7533201.

Wang, Q. and Chan, A. B. (2018). Cnn+ cnn: Convolutional decoders for image captioning. *arXiv preprint arXiv:1805.09019*. https://arxiv.org/pdf/1805.09019.

Wang, Y., Lin, Z., Shen, X., Cohen, S., and Cottrell, G. W. (2017). Skeleton key: Image captioning by skeleton-attribute decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_Skeleton_Key_Image_CVPR_2017_paper.pdf.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. https://doi.org/10.1007/978-3-319-46478-7_31.

Wu, Q., Shen, C., Wang, P., Dick, A., and Hengel, A. v. d. (2018a). Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1367–1381. https://doi.org/10.1109/TPAMI.2017.2708709.

Wu, X., He, R., Sun, Z., and Tan, T. (2018b). A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896. https://doi.org/10.1109/TIFS.2018.2833032.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning (ICML)*, pages 2048–2057. https://proceedings.mlr.press/v37/xuc15.pdf.

Yang, L., Tang, K., Yang, J., and Li, L.-J. (2016a). Dense captioning with joint inference and visual context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Yang_Dense_Captioning_With_CVPR_2017_paper.pdf.

Yang, L., Wang, H., Tang, P., and Li, Q. (2021). Captionnet: A tailor-made recurrent neural network for generating image descriptions. *IEEE Transactions on Multimedia*, 23:835–845. https://doi.org/10.1109/TMM.2020.2990074.

Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., and Salakhutdinov, R. R. (2016b). Review networks for caption generation. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9996535e07258a7bbfd8b132435c5962-Paper.pdf.

Yao, T., Pan, Y., Li, Y., and Mei, T. (2017a). Incorporating copying mechanism in image captioning for learning novel objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/papers/Yao_Incorporating_Copying_Mechanism_CVPR_2017_paper.pdf.

Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T. (2017b). Boosting image captioning with attributes. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4904–4912. https://openaccess.thecvf.com/content_ICCV_2017/papers/Yao_Boosting_Image_Captioning_ICCV_2017_paper.pdf.

Yao, X. (1999). Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447. https://doi.org/10.1109/5.784219.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2016/papers/You_Image_Captioning_With_CVPR_2016_paper.pdf.

Zhang, L., Sung, F., Liu, F., Xiang, T., Gong, S., Yang, Y., and Hospedales, T. M. (2017). Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*. https://www.robots.ox.ac.uk/~lz/AC_nips2017/ac_nips2017.pdf.

Zhang, Y., Lu, Y., Wu, H., Wen, C., and Ge, C. (2016). Face occlusion detection using cascaded convolutional neural network. In *Chinese Conference on Biometric Recognition*, pages 720–727. https://doi.org/10.1007/978-3-319-46654-5_79.

Zhang, Y., Shang, K., Wang, J., Li, N., and Zhang, M. (2018). Patch strategy for deep face recognition. *IET Image Processing*, 12(5):819–825. https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/iet-ipr.2017.1085.

Zhong, W. and Miyao, Y. (2021). Leveraging partial dependency trees to control image captions. In *Second Workshop on Advances in Language and Vision Research*, pages 16–21, Online. Association for Computational Linguistics. https://aclanthology.org/2021.alvr-1.3.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049. https://doi.org/10.1609/aaai.v34i07.7005.