1. INTRODUCCIÓN A LA VISIÓN POR COMPUTADOR

En este capítulo abordaremos dos temáticas. En la primera, se contextualiza la visión por computador dentro de la inteligencia artificial, así como se mencionan las principales empresas tecnológicas que investigan en este tema. En la segunda parte, se relacionan algunos de los trabajos más representativos de la visión por computador moderna.

VISIÓN POR COMPUTADOR EN LA ACTUALIDAD

Actualmente, gran parte de las compañías con mayor capitalización de mercado corresponden a empresas tecnológicas como Apple, Amazon, Microsoft, Alphabet, Facebook o NVIDIA. Estas empresas hacen parte del top 20 de los mayores proveedores de bienes y servicios a nivel mundial[1], por lo cual, sus temáticas de investigación son consideradas como tendencias en tecnología.

La Tabla I presenta un resumen de las áreas de investigación activas de algunas de estas empresas, en las que el procesamiento de señales (imagen, video, voz) y de datos, junto con la inteligencia artificial (aprendizaje de máquina, aprendizaje profundo), la visión por computador, y el procesamiento natural del lenguaje, son temáticas comunes entre ellas.

Tabla 1. Áreas de investigación activas en las mayores empresas tecnológicas actuales

Empresa	Algunas áreas de investigación activas
Apple[2]	Accesibility, Computer vision, Data science and annotation, Health, Privacy, Speech and natural language processing.

[1] https://companiesmarketcap.com/
[2] https://machinelearning.apple.com/research/

Amazon[3]	Cloud and systems, Computer vision, Natural language processing, Economics, Information and knowledge management, Machine learning, Operations research and optimization, Quantum technologies, Robotics.
Microsoft[4]	Artificial intelligence, Computer vision, Human computer interaction, Security, privacy and cryptography, Systems and networking.
Alphabet[5]	Algorithms and theory, Data mining and modeling, Information retrieval and the Web, Machine perception, Speech processing, Human-Computer Interaction and Visualization, Machine Intelligence, Natural Language Processing.
Facebook[6]	AR/VR, Artificial Intelligence, Computer vision, Data science, Databases, Machine learning, Natural Language Processing & speech, Security & Privacy.
NVIDIA[7]	3D, Deep learning, Artificial intelligence and machine learning, Computational photography and imaging, Computer graphics, Computer vision, Real-time rendering, VR, AR and display technology.

En el caso particular de la visión por computador, cabe resaltar que este es un campo interdisciplinario que puede involucrarse con diferentes áreas de la ciencia, ingeniería o tecnología. Por ejemplo, en la física, donde se estudia la óptica y la formación física de las imágenes; o en las ciencias biológicas y psicológicas, para entender cómo se procesa físicamente la información visual. Con las matemáticas e ingeniería de software, para el avance en la implementación de algoritmos de visión por computador.

En la más reciente década los principales avances en visión por computador se han alcanzado gracias a la aplicación de técnicas de inteligencia artificial, inicialmente con esquemas de aprendizaje de máquina y luego, específicamente,

[3] https://www.amazon.science/
[4] https://www.microsoft.com/en-us/research/
[5] https://research.google/research-areas/
[6] https://research.fb.com/research-areas/
[7] https://www.nvidia.com/en-us/research/

con el uso de herramientas de aprendizaje profundo. Es importante señalar que las técnicas de aprendizaje profundo son un subconjunto de técnicas de aprendizaje de máquina, y estas a su vez son un subconjunto de técnicas de inteligencia artificial, como lo muestra la Figura 1. Sin embargo, aunque se han presentado avances importantes en los últimos años, la inteligencia artificial no es un concepto nuevo, y sus principales desarrollos se han venido postulando desde mediados del siglo pasado.

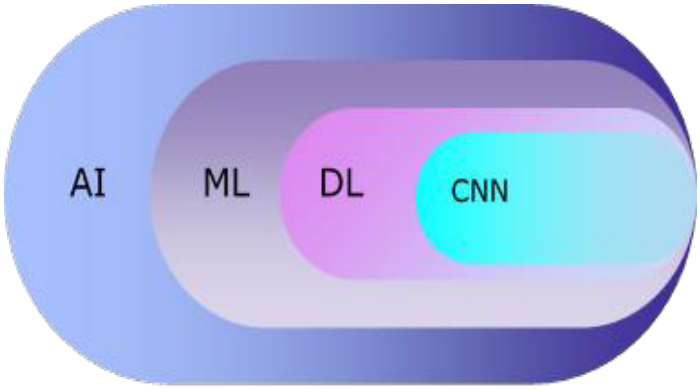

Figura 1. CNN y aprendizaje profundo en el contexto de aprendizaje automático e inteligencia artificial

Como contexto al aporte de la IA, se puede decir que la inteligencia artificial se define como cualquier técnica que permite a los computadores imitar el comportamiento humano. Por su parte, el aprendizaje de máquina permite alcanzar cierto grado de inteligencia artificial en sistemas que pueden aprender de la experiencia para reconocer patrones en un conjunto de datos. Por ejemplo, los sistemas basados en aprendizaje profundo son un caso específico de aprendizaje estadístico para extraer características o atributos de un conjunto de datos en bruto mediante el uso de múltiples capas ocultas, mientras que, los esquemas de aprendizaje de máquina, típicamente requieren que la extracción de características se realiza de forma manual, como etapa previa al modelamiento.

Por otra parte, es posible discriminar diferentes tipos de aprendizaje profundo: aprendizaje supervisado, aprendizaje no supervisado o aprendizaje por refuerzo. El aprendizaje supervisado se caracteriza principalmente por contar con un conjunto de datos que involucra tanto las características de los datos que serán procesados (datos de entrada), así como una etiqueta que los define (lo que tiene que aprender el algoritmo). En este caso, se habla de un conjunto de datos que permitirá tanto entrenar un algoritmo, como evaluarlo con datos correctamente etiquetados (*ground truth*). Como ejemplos de modelos que utilizan este tipo de

aprendizaje se tienen las redes neuronales convolucionales, las redes neuronales recurrentes o las arquitecturas tipo codificador-decodificador (*encoder-decoder*).

Por el contrario, en el caso de aprendizaje no supervisado, no se cuenta con un conjunto de datos etiquetados, por lo cual el algoritmo deberá identificar por su cuenta los patrones que caracterizan a los datos sin una referencia específica (*ground truth*). Un ejemplo de ello, son los algoritmos de clusterización. Por su parte, el aprendizaje por refuerzo involucra modelos que a partir de la representación de la información aprenden a realizar acciones a partir de una especie de realimentación (recompensas o penalizaciones), lo que a su vez puede modificar el estado del entorno8.

En cualquier caso, el aprendizaje deberá realizarse a partir de las características o atributos de los datos de entrada. Un ejemplo sencillo de atributos puede ser un clasificador de imágenes de frutas, en el cual las entradas son el peso y el color. Con un conjunto de datos reducido, esto permitirá distinguir algunas frutas entre sí, pero tal vez no será suficiente para generalizar el problema. En un contexto real, es probable que el conjunto de datos de entrada sea de una dimensionalidad mucho más alta y que a su vez el problema tenga un mayor número de datos. De hecho, algunos conjuntos de datos de reconocimiento de imágenes cuentan con millones de imágenes. Por esta razón, es importante tener en cuenta el tamaño del conjunto de datos, las entradas o características del modelo, la disponibilidad de los datos, al momento de trabajar en un proyecto de ciencia de datos.

Partiendo de lo anterior, se resaltan dos aspectos clave que permitieron la evolución del aprendizaje automático hacia el aprendizaje profundo. El primero, tiene que ver con una mayor cantidad de datos disponibles, propiciado por la masificación de tecnologías como Internet o la telefonía móvil, y sus respectivas aplicaciones. El segundo aspecto tiene que ver con las mejoras en el rendimiento de sistemas computacionales, por ejemplo, a partir de computación paralela o la disponibilidad a mayor escala de unidades de procesamiento gráfico (GPU). Estos dos aspectos permitieron no solo alcanzar un mayor rendimiento, si no también que este rendimiento se alcanzará a través de redes neuronales más profundas, tal como lo ilustra la Figura 2 (Saravia, 2021).

[8] https://deeplearning.mit.edu/

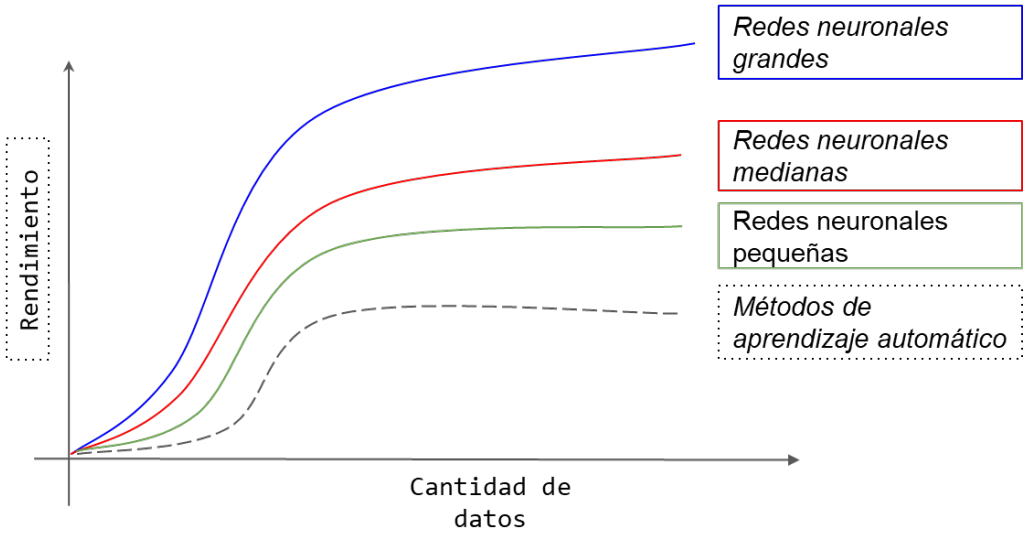

Figura 2. Rendimiento versus tamaño de un conjunto de datos en modelos de aprendizaje automático.

Respecto a los principales tipos de problemas, o las salidas que puede entregar un algoritmo de aprendizaje automático se tienen la regresión, clasificación, *clustering* o predicción de secuencia. La regresión se caracteriza por predecir un valor continuo, por ejemplo, predecir el valor de un bien inmueble a partir de sus características (antigüedad, tamaño, sector, etc.). Los modelos de clasificación se encargan de predecir un valor discreto (o categoría), por ejemplo, en la predicción de la raza de un perro a partir de una imagen. Por su parte, la función de los algoritmos de *clustering* consiste en realizar la segmentación o pertenencia a un grupo de casos similares, por ejemplo, en la segmentación de clientes de un supermercado a partir de su historial de compra. En cuanto a la predicción de secuencias, estas se encargan de estimar los valores siguientes de la misma, por ejemplo, en los sistemas de texto predictivo presentes en los dispositivos móviles.

TRABAJOS REPRESENTATIVOS EN VISIÓN POR COMPUTADOR

Aunque el auge actual de modelos y algoritmos basados en visión por computador ha presentado enormes y valiosos aportes en la última década, su desarrollo se ha venido construyendo décadas atrás. De hecho, las redes neuronales aparecen por primera vez a finales de la década del 50 del siglo pasado, en un trabajo donde Frank Rosenblatt del *Cornell Aeronautical Laboratory* propone el uso del perceptrón de una sola capa (Rosenblatt, 1958). Esta arquitectura fue posteriormente generalizada al conocido perceptrón multicapa (MLP), el cual

sigue siendo inspiración para actuales desarrollos (Liu, 2021), (Tolstikhin, 2021). Sin embargo, un aporte fundamental para el uso de esta arquitectura fue el planteamiento del proceso de aprendizaje por retro propagación (*backpropagation*) para redes neuronales, el cual ajusta de manera iterativa los pesos de las conexiones de la red para minimizar la diferencia entre el valor estimado por la red y su etiqueta o valor real (Rumelhart, 1986). Un trabajo posterior va un paso más allá, aplicando el algoritmo de retro propagación al reconocimiento de dígitos escritos a mano en documentos postales (LeCun Y. B., 1989), misma temática trabajada en la primera red neuronal convolucional propuesta por el mismo autor (LeCun Y. B., 1998).

Como un primer acercamiento para la comprensión de la visión a nivel biológico, el trabajo presentado por (Hubel, 1962) y realizado sobre gatos aplicó electrofisiología para estimular la corteza visual primaria mediante filtros orientados y describir así su respuesta. Dichos autores encontraron que las células simples responden a la orientación de la luz, las células complejas responden tanto a la orientación de luz como al movimiento y que células complejas responden a movimientos de luz en puntos específicos (Hubel, 1962).

Para la siguiente década, se presentaron dos iniciativas en el Instituto de Tecnología de Massachusetts (MIT). Una de ellas se considera la primera tesis de doctorado en visión por computador, y estuvo orientada hacia el modelado de sólidos en 3D para el reconocimiento y reconstrucción de formas geométricas simples del mundo real (Roberts, 1963). La segunda fue un proyecto de verano orientado a la solución de tareas concretas en visión por computador para el reconocimiento de patrones (Papert, 1966).

En un trabajo posterior desarrollado en este mismo instituto en la década de los 70, y publicado de manera póstuma (Marr D. , 2010), el autor describe un marco general para entender la percepción visual, abordando cuestiones sobre el estudio y comprensión del cerebro y sus funciones para la construcción de representaciones a partir de la descripción de una imagen y la descripción de objetos tridimensionales del entorno. Un aspecto importante es la introducción por parte de dicho autor sobre nociones para diferentes niveles de análisis: nivel computacional, nivel algorítmico y nivel de implementación de hardware.

De forma paralela, se abordó un problema fundamental en la representación de objetos: reducir la estructura compleja de un objeto a una colección de formas más simples y su configuración geométrica. En este sentido, se desarrollaron trabajos para la representación del cuerpo humano, obteniendo representaciones

mediante cilindros (Brooks, 1979) y mediante una estructura pictórica (Fischler, 1973). Estos trabajos constituyen la base de los modelos utilizados hoy en día.

Posteriormente se presentan trabajos enfocados en el reconocimiento de bordes y líneas, así como la segmentación de áreas de la imagen. Aquí se tienen trabajos pioneros en el área, como el reconocimiento de objetos 3D a partir de imágenes en 2D (Lowe D. G., 1987), la segmentación binaria de imágenes para resolver el problema de agrupación perceptiva en la visión (Shi, 2000), el reconocimiento de objetos a partir de detección de puntos clave mediante características SIFT (*Scale-Invariant Feature Transform*) (Lowe D. , 1999). También se tienen trabajos relacionados con el reconocimiento de categorías de escenas mediante la división de la imagen en subregiones y el correspondiente cálculo de histogramas de las características locales de cada subregión obteniendo una representación llamada pirámide espacial (Lazebnik, 2006). Otros importantes trabajos en esta temática involucran fundamentos de detección de bordes (Marr D. &., 1980), o algoritmos fundamentales y ampliamente conocidos para detección de bordes (Canny, 1986), segmentación de imágenes basada en regiones (Chan, 2001) y detección de ángulos (Harris, 1988).

También se tienen trabajos para la detección de personas o partes de estas (como el reconocimiento de rostros). Por ejemplo, los autores en (Dalal, 2005), proponen el uso de descriptores de gradientes orientados (HOG) para el reconocimiento robusto de objetos visuales (particularmente detección humana basada en SVM lineal). Por su parte, los autores en (Felzenszwalb, 2008), proponen la detección de personas mediante un modelo de detección de objetos entrenado para reconocer diferentes formas a diferentes escalas. Respecto a la detección de rostros, uno de los trabajos más relevantes está basado en la transformada wavelet con filtro Haar y un clasificador Adaboost (Viola, 2004).

Más allá de estos importantes aportes, uno de los contextos que impulsó en gran medida el desarrollo de nuevas soluciones en el área de visión por computador fue el establecimiento de desafíos abiertos a la comunidad científica. De forma particular, el desafío PASCAL para reconocimiento de objetos visuales, establecido con veinte categorías de objetos y alrededor de diez mil imágenes por clase, y el desafío de clasificación de imágenes, IMAGENET, establecido inicialmente con mil clases y mil imágenes por clase (1 millón de imágenes), se convirtieron en referentes para el desarrollo de nuevos modelos para reconocimiento de objetos, como por ejemplo nuevas arquitecturas de aprendizaje profundo.

Uno de los aspectos a resaltar aquí, radica en que el desafío IMAGENET permitió el surgimiento de las redes neuronales convolucionales, las cuales alcanzaron un rendimiento superior sobre las técnicas de visión por computador tradicionales, lo que llevaría al auge del aprendizaje profundo. Este punto de quiebre se presentó en la edición 2012 de este desafío, con una arquitectura de 8 capas de profundidad conocida como AlexNet (Krizhevsky, 2012). Para el año siguiente el primer puesto del desafío lo obtiene igualmente una arquitectura de 8 capas de profundidad, para la cual los autores exploraron técnicas de visualización que permitieron conocer la función de las capas de características intermedias y el funcionamiento del clasificador, superando el rendimiento de la arquitectura AlexNet (Zeiler, 2014).

Para el año 2014, surgen dos arquitecturas fundamentales en el desarrollo del aprendizaje profundo: VGG (Simonyan, 2014), una arquitectura de 19 capas para el reconocimiento de imágenes a gran escala y GoogLeNet (Szegedy, 2015), una arquitectura de 22 capas cuyo rendimiento fue superior al alcanzado por arquitecturas previas. Otra arquitectura fundamental en el estado del arte es ResNet, caracterizada por su gran profundidad (152 capas) y que se presentó en 2015, siendo premiada como la mejor ponencia en la conferencia sobre visión por computador y reconocimiento de patrones (CVPR) (He, Deep residual learning for image recognition, 2016).

La investigación y aportes relacionados hasta aquí, sumados a una gran cantidad de desarrollo y propuestas adicionales realizados por diferentes empresas, universidades e investigadores han permitido el avance de esta área, mostrando un avance significativo en la última década. Esto ha llevado a que el sector gubernamental, las universidades y los departamentos de investigación y desarrollo de las grandes empresas de tecnología hayan llevado su atención hacia temáticas relacionadas con este tema. Dada su importancia, esta es precisamente el área que se tratará en el presente libro.

