GOOGLE LENS

Trabajo de innovación

Marc Nebot
Victor Teixidó
Florian Vogel

ÍNDICE

1. Introducción

En este trabajo vamos a profundizar sobre el funcionamiento de uno de los proyectos innovadores de *Google*, *Google Lens*. Fue lanzado en 2017 y, a pesar de que al principio no llamó mucho la atención, poco a poco se ha ido integrando de forma natural e imperceptible en nuestro día a día. Este proyecto representa un claro ejemplo de cómo la inteligencia artificial puede simplificar tareas complicadas como traducir un texto. Con la herramienta se pueden realizar búsquedas visuales a través de la cámara de un smartphone en tiempo real.

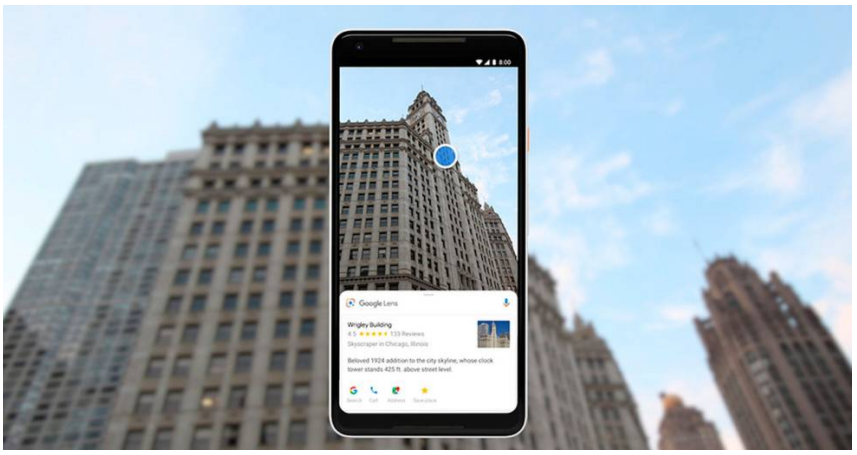
En el primer apartado hablaremos de *Google Lens*, sus funcionalidades y utilidades y mostraremos de lo que llega a ser capaz. Siendo un software tan amplio y complejo, hemos decidido centrarnos en el apartado de la lectura y conversión de texto que puede realizar *Google Lens* aunque hablaremos también de la búsqueda a partir de imágenes también. Veremos y explicaremos las técnicas de inteligencia artificial utilizadas por *Google* para que esto sea posible. Por último, concluiremos con el impacto real de una herramienta como esta en la actualidad y en la vida cotidiana de las personas, además de ver el potencial de cara al futuro que podría llegar tener.

2. Google Lens

Google Lens es una tecnología desarrollada por *Google* y lanzada inicialmente en 2017. A través del reconocimiento de imágenes, *Google Lens* trata de obtener información relevante de los objetos que el usuario escanea a través de la cámara del smartphone.

El usuario puede apuntar la cámara a cualquier objeto y la Lens-App intenta desarrollar una comprensión del objeto para luego mostrar diferentes informaciones sobre el resultado al usuario.

La imagen de la derecha muestra el aspecto de la interfaz de usuario básica al utilizar la aplicación.



Img 1: Lens UI

Sus casos de uso son muy variados. Por ejemplo, puedes buscar simplemente la ropa que ves y quieres saber dónde comprarla. O tal vez en un país extranjero quieras traducir la leyenda de un alimento. En un clip publicitario, *Google* demuestra que la función de texto a voz, integrada en *Google Lens*, puede ayudar a las personas que tienen dificultades para leer a entender el mundo que les rodea.

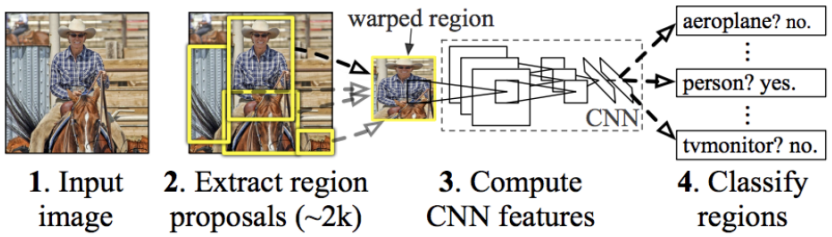
La herramienta tiene muchas otras funciones integradas que, en conjunto, dan lugar a un gran proyecto que quizá no está terminado, ya que no todo funciona a la perfección, pero demuestra el gran potencial que tiene la IA para simplificar nuestras vidas y que ya puede ser de gran ayuda en muchas situaciones.

3. Uso de técnicas de IA

En esta sección hablaremos de cómo *Google* hizo posible las funcionalidades de *Google Lens* y qué técnicas concretas se utilizaron en relación con la IA. Más adelante, en la cuarta sección, explicaremos cómo funcionan algunas técnicas fundamentales que potencian Lens. La aplicación tiene grandes ambiciones. *Google* quiere que pueda funcionar en una amplia gama de dispositivos. Esto es especialmente importante, ya que existe un gran mercado en los países en desarrollo y muchas personas analfabetas que viven allí. Para esas personas el proyecto representa una solución para integrarse mejor en la vida cotidiana (véase la introducción). Para poder ofrecer el rendimiento necesario para realizar las tareas solicitadas, incluso en dispositivos de gama baja, la clave son los algoritmos eficientes. Ahora vamos a señalar qué algoritmos utiliza *Google Lens* para implementar sus funciones.

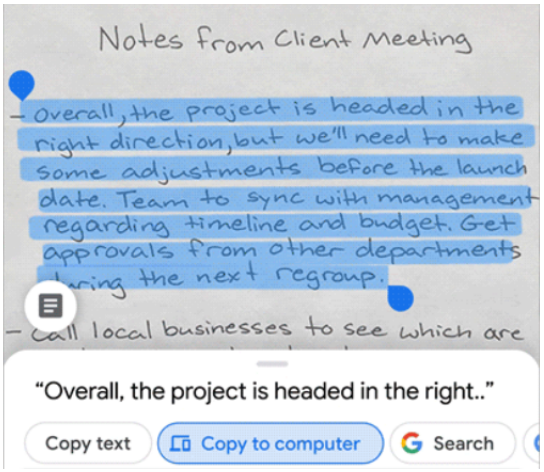
Convolutional Neural Network (CNN)

Cuando el software de *Google Lens* recibe una imagen, el primer paso a realizar es tratar la imagen para discernir los diferentes objetos que se puedan encontrar en esta. Para realizar esto, se utilizan algoritmos CNN, Convolutional Neural Network. El funcionamiento de estas redes neuronales es el siguiente. Inicialmente se seleccionan 2000 regiones dentro de la imagen que serán los posibles objetos, una vez se tienen todas estas regiones, haciendo uso de algoritmos greedy, se irán fusionando recursivamente aquellas regiones que formen parte de un objeto más grande. Haciendo esto, los candidatos finales que queden, serán las regiones evaluadas por los sistemas entrenados de identificación de objetos. Un problema que tienen los CNN es que si quisiéramos analizar peatones que se encuentran muy lejos en la foto, necesitaríamos de técnicas adicionales para su detección. La identificación de un objeto y el aprendizaje están explicados en mayor profundidad en el apartado 4.



Una vez hemos identificado de qué objeto se trata, *Google Lens* mostrará información al respecto. Si por ejemplo el objeto es una planta, se enseñará en pantalla de que tipo es, imagenes… O si por ejemplo fuera una pieza de ropa, se mostrarían tiendas en las que comprar la misma, o similares, piezas de ropa. Entre estos casos, cabe destacar cuando estos objetos son textos, ya que el software de *Google Lens* requerirá de hacer algunos pasos adicionales para su manipulación y tratamiento.

Optical Character Recognition (OCR)



El reconocimiento de carácteres, o en inglés *Optical Character Recognition*, describe el proceso de identificación de carácteres en textos impresos o escritos a manos y la conversión de estos a un formato digital. El objetivo de esto, es buscar la conversión de textos no digitales y acabar obteniendo datos procesados que puedan ser utilizados por softwares. El *Optical Character Recognition* (OCR), es un área bastante importante dentro de la visión de computadores y, en los últimos años, está teniendo tiene un gran importancia dentro de campos relacionados con la inteligencia artificial y tecnologías que hacen uso de esta, un claro ejemplo como estamos viendo es el software de *Google Lens* que hace acopio de estas herramientas.

Pese a que el concepto y el objetivo del reconocimiento de textos sea claro y sencillo, el funcionamiento y obtención de estos puede convertirse en una tarea altamente complicada cuando hablamos de textos escritos a mano, debido a la infinitud de variables y diferencias que puedan presentarse. Aún así, vamos a resumir el proceso que realiza *OCR* en 3 pasos: Preprocesado de la imagen, reconocimiento de carácteres y postprocesado.

Preprocesado

Este paso previo a la identificación del texto en una imagen consiste en la obtención de la imagen y varios procesos de facilitación de lectura. Una vez obtenida la imagen, se eliminan imperfecciones que puedan complicar la lectura e identificación de caracteres, así como se suavizan los píxeles buscando obtener una imagen más clara y limpia. Por último, se hace una binarización de la imagen respecto al fondo y los carácteres, donde el fondo estará representado por zonas claras y el texto por zonas oscuras, esto facilitará más adelante la criba entre los carácteres. Este paso irá unido con el algoritmo de CNN de detección de objetos explicado anteriormente.

Reconocimiento de carácteres

Las zonas oscuras de la imagen resultante, el texto, serán tratadas para identificar los diferentes carácteres. Uno de los algoritmos utilizados para esto es el de detección de características, es decir el número de esquinas, curvas, cruces, líneas rectas… que hay en cada uno de los carácteres. Algoritmos para la detección de características hay varios pero todos están basados en la misma idea ya mencionada, mirar los puntos de interés similares y aquello que hay a su alrededor, y en base a esto, otorgar puntos de similitud a cada uno de los símbolos respecto a los carácteres. Algoritmos de OCR avanzados dividen cada símbolo o letra en los diferentes componentes mencionados y los compará y relaciona con partes y secciones de carácteres existentes.

Una vez se ha obtenido el resultado de los carácteres y posibles frases o textos que estos formen, se le dará al usuario la posibilidad de traducir estos textos, copiarlos, escucharlos en voz alta… Esta última funcionalidad la veremos en el siguiente punto con más detalle y profundidad.

Text-To-Speech

Tras la extracción del texto detectado *Google Lens* ofrece una funcionalidad de texto a voz (TTS) que lee el texto en voz alta. Por supuesto, esta no es la única aplicación de *Google* en la que se utiliza esta función, sin embargo, queremos echar un vistazo más de cerca, ya que podría ser la funcionalidad más utilizada, especialmente para las personas que no saben leer.

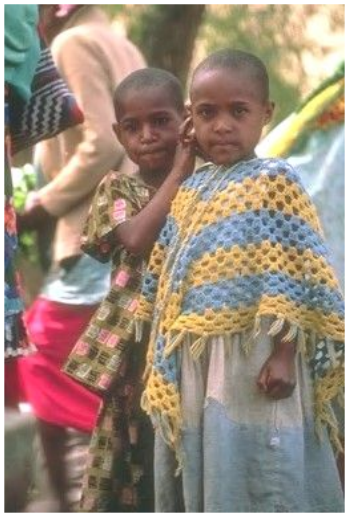
En general, hay dos enfoques para generar voz sintetizada. O bien concatenando trozos de habla grabada (TTS concatenativo) o creando una salida completamente "sintética".
Este último enfoque tiene algunas ventajas importantes, por ejemplo, es más fácil modificar la voz o expresar diferentes emociones sin tener que grabar toda una nueva base de datos. Hasta hace unos años, para generar esta salida de audio completamente "sintética" se utilizaba una técnica llamada TTS paramétrica en combinación con vocoders. Con esta técnica es fácil alterar las características del habla, pero en TTS paramétrico ha tendido a sonar menos natural que el concatémero.

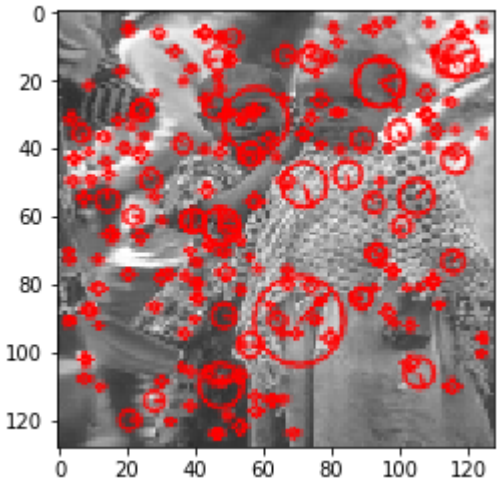
En 2016 DeepMind (*Google*) lanzó una nueva técnica, para generar una salida puramente "sintética", llamada WaveNet. Aporta importantes mejoras en términos de inteligibilidad y naturalidad de la voz. La técnica permite cambiar la identidad del hablante, expresar emociones o imitar acentos. Aunque al principio no era lo suficientemente eficaz como para funcionar en teléfonos inteligentes, en los últimos años ha recibido muchas mejoras y ahora impulsa la función de texto a voz de *Google*. WaveNet se basa en redes neuronales convolucionales, una modificación de las redes neuronales profundas de la que hablaremos con más detalle en la sección cuatro.

Content Based

Otra de las técnicas de *IA* que aplica *Google Lens* es el *Content Based*. El *Content Based* es una técnica que se suele emplear en sistemas de recomendación como por ejemplo películas pero esta vez será utilizado en imágenes. En este caso, *Google Lens* aplica *Content Based* para recomendar imágenes similares a la que el usuario detecta con la cámara siguiendo una de las técnicas de detección de imágenes que ya hemos mencionado antes. El procedimiento será el siguiente.

Por ejemplo tenemos esta imagen junto a su escaneo de *Google Lens*:





Lo que hará el sistema, en vez de asociar texto a la imagen, será detectar el color, la textura, la forma, la posición y quizá alguna característica más de cada uno de los píxeles, los parametriza y buscará las imágenes con el mayor índice de similaridad a las características parametrizadas de la imagen que el usuario le ha introducido. Una vez sean obtenidas las imágenes más similares a la imagen fotografiada por el usuario o simplemente subida de la galería, *Google Lens* mostrará por pantalla las *k* imágenes más similares.

Por este motivo podemos afirmar que *Google Lens* utiliza *Content Based* a la hora de encontrar las imágenes más similares a un modelo dado.

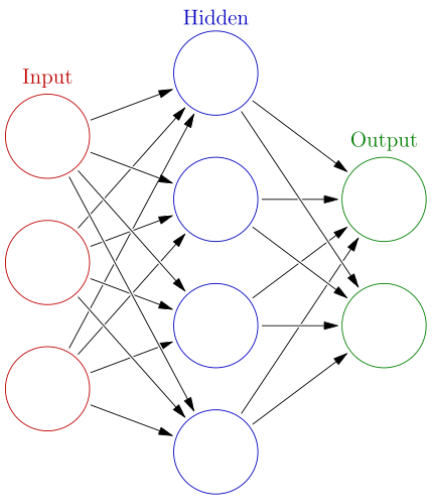
4. Técnicas en profundidad

4.1 Redes neuronales artificiales (RNA)

Las RNA son una técnica de aprendizaje automático que se inspira en las redes neuronales biológicas del cerebro. Sus principales componentes son los nodos (llamados neuronas) que simulan las neuronas biológicas y las aristas que simulan las sinapsis biológicas.

Las neuronas tienen un valor que se calcula mediante la suma de todos los valores de las aristas entrantes y una función específica. El valor puede ser transmitido a otras neuronas de la capa siguiente (la neurona produce una única salida, que puede ser transmitida a varias neuronas) mediante las aristas.

Las neuronas de las RNA se distribuyen en tres tipos de capas diferentes:



Img 5: ANN capas

Una capa de entrada, una o varias capas ocultas y una capa de salida, la señal suele fluir de izquierda a derecha. La capa de entrada recibe la información externa. La capa de salida produce el resultado final.

Mecanismo

Para que la red obtenga un resultado para una entrada determinada, primero hay que pasar los valores de entrada a la primera capa. A continuación, para cada capa (una tras otra, de izquierda a derecha) cada neurona calcula su valor de salida. Primero calculamos la activación de la neurona calculando la suma de todas las entradas, ponderada por un peso y añadiendo un sesgo.

Este valor se pasa a una función de activación que calcula la salida. A menudo se trata de una función no lineal, por ejemplo la sigma σ (que tendrá un valor entre 0 y 1).
De forma un poco más técnica, la función de activación podría definirse como

$$\frac{1}{1+\exp(-\sum_j w_j x_j - b)}$$

Img 6: Función de activación

Siendo x1, x2,... las entradas, w1, w2,... los pesos correspondientes y b el sesgo.

Las entradas son los valores de las neuronas con aristas que conducen a la neurona en cuestión, o si la neurona se encuentra en la primera capa la entrada a la neurona es la entrada al gráfico.

El peso describe la importancia de la arista, para determinar su valor específico la red necesita someterse a un proceso de aprendizaje que describiremos a continuación en el apartado de aprendizaje.

El sesgo representa una constante añadida a la entrada, con la que se puede desplazar el valor de la función de activación. También el valor exacto del sesgo es el resultado del proceso de aprendizaje.

Como ya se ha mencionado, la salida de la red se obtiene mirando los valores de la función de activación en la última capa. Normalmente obtenemos un valor para cada neurona en el rango de cero a uno. Si buscamos un resultado, el nodo con el valor más alto lo representa.

Aprendizaje

Ahora que conocemos la estructura general de la RNA, podemos ver cómo una red de este tipo puede realizar un aprendizaje, por ejemplo, clasificar imágenes.

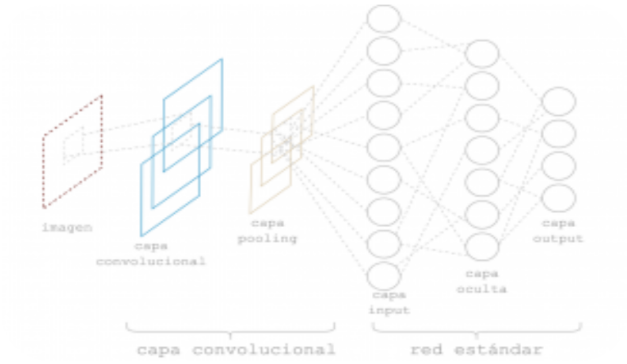
En primer lugar, necesitamos un conjunto de datos de entrenamiento (aprendizaje supervisado), que servirá de entrada a la red. Durante el proceso de aprendizaje, la red intenta ajustar los pesos y los sesgos de todas las neuronas y aristas para minimizar el error observado. Este error representa la diferencia entre la salida de la red y la respuesta correcta.

Para calcular el error tenemos que definir una función de coste. Esta función cuantifica el rendimiento de la red. Un ejemplo típico sería el error medio al cuadrado (MSE), aunque esta función es una parte crucial a la hora de modificar la red para aumentar su rendimiento.

El objetivo entonces es encontrar pesos y sesgos que hagan que el coste sea lo más pequeño posible. Para esta tarea se suele utilizar el algoritmo de descenso de gradiente o alguna variación del mismo.

4.2 Redes Neuronales Convolucionales

Hemos visto que *Google Lens*, y otras herramientas que trabajen con detección de imágenes, utilizan redes neuronales convolucionales. Estas no son más que una ampliación a las redes neuronales estándar, o redes neuronales artificiales, que ya hemos visto en el apartado anterior. La ampliación que precede a este tipo de redes consiste en añadir dos capas adicionales al principio de la red, que tendrán el propósito de manipular los píxeles de las imágenes.



Como se puede apreciar en la imagen anterior, esta ampliación estará formada por dos capas adicionales: la capa convolucional y la capa de pooling.

La capa convolucional procesa los píxeles cercanos a la entrada, algo así como los píxeles de una región local, y calcula el producto escalar entre los valores de píxel y una pequeña zona a la que están conectados en el volumen de entrada. Por otra parte, se encarga de extraer los píxeles más representativos de una zona delimitada de una imagen. Además de estas dos capas adicionales pre-red neuronal, el resto del funcionamiento de las redes neuronales convolucionales, será igual a las redes neuronales estándar.

5. Innovación e Impacto

Aspectos Innovadores del producto

Google Lens en sí no inventa ninguna técnica nueva y no es una idea revolucionaria. Lens utiliza el estándar actual en lo que respecta a muchas funciones, como el reconocimiento de imágenes (por ejemplo, OCR), el procesamiento del lenguaje natural (por ejemplo, texto a voz) o los gráficos de conocimiento (por ejemplo, el gráfico de conocimiento de Google).

En el proceso, Google ha desarrollado la tecnología líder para el reconocimiento de imágenes. Además, Google ha realizado varios avances en el aprendizaje automático para hacer posible este producto (por ejemplo, Wavenet).

Por lo tanto, creemos que Lens como producto representa lo que los equipos de investigación de Google han logrado en los últimos años y que Google es un importante pionero en el campo de la IA.

Impacto en la empresa

Beneficios

Uno de los principales beneficios que una herramienta como *Google Lens* aporta a *Google* es ser la primera empresa en haber conseguido unificar en una única aplicación, todo lo que *Google Lens* hace. Puedes desde traducir a tiempo real texto, identificar tipos de animales o plantas, detectar ropa o muebles, leer códigos QR o incluso hacer una foto a una tarjeta de presentación y poder añadir un contacto directamente. Esto otorga un prestigio notorio a la empresa y le permite seguir siendo líder en innovación y tecnología como lo es actualmente.

Riesgos

El principal riesgo de un proyecto como *Google Lens*, y como cualquier otro, es su total y absoluto fracaso y, por tanto, el malgasto de recursos, horas, dinero y personal invertido en él. Ahora bien, hablando de una empresa como es *Google*, estos riesgos están altamente minimizados ya que pueden permitirse el lujo, dada la posición en la que se encuentran, de probar proyectos más ambiciosos sin miedo a que estos puedan llegar a fracasar.

Posición de mercado

La posición de mercado de *Google Lens*, actualmente es muy buena. En este sentido, no existen otros competidores reales que puedan generar competencia a *Google*, además, el simple hecho de que venga por defecto instalado en la propia cámara de algunos teléfonos, hace que sea aún más difícil, por no decir imposible, competir contra este producto. ¿Por qué querrías instalar una aplicación adicional que haga lo mismo que y hace tu cámara? Hay que destacar que en este sentido, *Google* tiene una clara ventaja competitiva respecto a los posibles competidores que puedan surgir.

Impacto en los usuarios

Beneficios

Google Lens no tuvo mucho impacto en cuanto salió pero este se fue acomodando poco a poco y actualmente, ha facilitado el dia a dia de los usuarios desde ayudándolos a encontrar una prenda de vestir que ha aparecido en la televisión y no saben de dónde es hasta el escaneo de un texto escrito a mano y pasarlo a digital.

A parte de ayudar en el día a día de los usuarios, esta herramienta ha sido utilizada también con fines educativos en distintas instituciones donde daban a alumnos dispositivos móviles compatibles con *Google Lens* y dejaban que analizasen las distintas plantas de un jardín para aprender más sobre ellas. Este mismo método también ha sido utilizado en agencias turísticas, permitían a los turistas examinar edificios con esta herramienta para obtener información de ellos. Podemos observar que a parte del uso diario, esta herramienta ha fomentado el autoaprendizaje en distintas áreas.

Riesgos

A pesar de las múltiples ventajas que *Google Lens* tiene en el día a día de los usuarios, este posee unas grandes desventajas. *Lens* en la actualidad, no es compatible todavía con muchos dispositivos por lo tanto, podemos observar que a pesar de tener un gran potencial como herramienta de uso diario, sin *Internet* ni un buen dispositivo móvil no podremos disponer de *Google Lens.*

Esta aplicación todavía está en desarrollo por lo tanto, tiene un alto índice de fallo cosa que lo hace poco eficiente y deja descontento al usuario, esto no quita que tenga un gran potencial y en un futuro no muy lejano el índice de fallo pueda ser reducido a casi cero.

6. Bibliografía

*[**1**] Rohith Gandhi. (2018, 9 julio). Object Detection Algorithms*

https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e

*[**2**] Google Lens: Urmila's Story. (2019, 7 mayo). [Vídeo]. YouTube.*

https://www.youtube.com/watch?v=ePwKgKp69GE

*[**3**] Wikipedia contributors. (2021, 7 febrero). Region Based Convolutional Neural Networks. Wikipedia. https://en.wikipedia.org/wiki/Region_Based_Convolutional_Neural_Networks*

*[**4**] Wavenet-launches-google-assistant. (2016). Wavenet-launches-google-assistant.*

https://deepmind.com/blog/article/wavenet-launches-google-assistant

*[**5**] Wavenet-generative-model-raw-audio. (2016). Wavenet-generative-model-raw-audio.*

https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

*[**6**] High-fidelity-speech-synthesis-wavenet. (2016). High-fidelity-speech-synthesis-wavenet.*

https://deepmind.com/blog/article/high-fidelity-speech-synthesis-wavenet

*[**7**] WAVENET: A GENERATIVE MODEL FOR RAW AUDIO. (2015).*

https://arxiv.org/pdf/1609.03499.pdf

*[**8**] Sagar, R. (2020, 4 agosto). These Machine Learning Techniques Make Google Lens A Success. Analytics India Magazine.*

https://analyticsindiamag.com/these-machine-learning-techniques-make-google-lens-a-success/

*[**9**] Naik, A. R. (2021, 7 octubre). Deep Learning For Computer Vision: A Brief History and Key Trends. Analytics India Magazine.* https://analyticsindiamag.com/deep-learning-for-computer-vision-a-brief-history-and-key-trends/

*[**10**] Das, R. (2021, 15 febrero). Why Google Lens Is The Perfect Answer For Content Based Image Recognition. Analytics India Magazine.* https://analyticsindiamag.com/why-google-lens-is-the-perfect-answer-for-content-based-image-recognition/

*[**11**] Techbyte-2019 | Insight. (2019). Techbyte-2019 | Insight.* https://www.jimsindia.org/techbyte2k19/insight/Deep%20Learning%20Architecture%20And%20Algorithms.html

*[**12**] Boesch, G. (2021, 27 junio). Optical Character Recognition (OCR) – Overview and Use Cases. Viso.Ai.* https://viso.ai/computer-vision/optical-character-recognition-ocr/

*[**13**] Verde, A. (2019, 10 julio). ¿ Cuantas veces has visto una bonita planta y has dicho.. ¡ Read more. El Poder Del AnDrOiDe VeRdE.* https://elpoderdelandroideverde.com/google-lens-que-es-y-para-que-sirve/

*[**14**] Ojeda, A., Gómez, S., Malik, A. M., & Gómez, S. (2015). Nebulova - Predicción con redes neuronales convolucionales (CNN). Nebulova - Predicción con redes neuronales convolucionales (CNN).* https://www.nebulova.es/blog/redes-neuronales-convolucionales

*[**15**] Juan Barrios. (2019, 18 junio). Redes neuronales convolucionales son un tipo de redes neuronales.* https://www.juanbarrios.com/redes-neurales-convolucionales/