GOOGLE LENS

Trabajo de innovación

Marc Nebot

Víctor Teixidó

Florian Vogel

ÍNDICE

1. Introducción

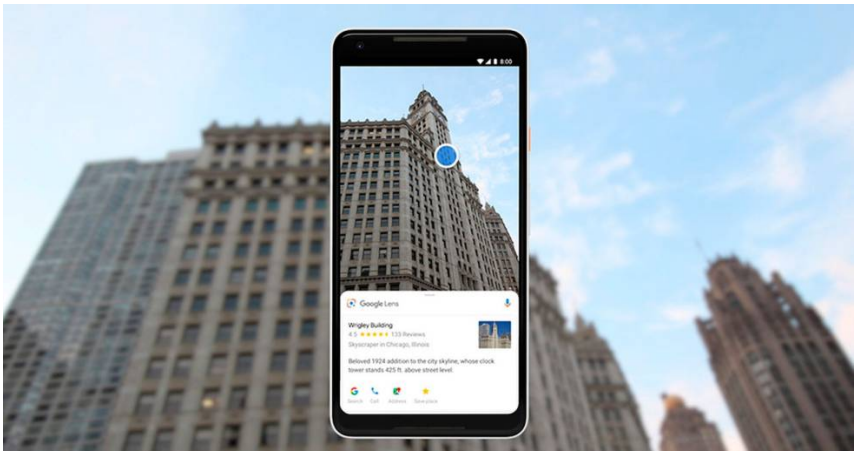
In this work we will take a deep look at one of Google's innovative projects. Google Lens was released in 2017 and though it didn't attract much public attention we think it might be one of the best examples of how AI can simplify many everyday tasks.
With the tool you can perform a visual search in real time through your smartphone camera and the data google has on you.

First we will Present the Google Lens project in general and show what it is capable of.
Then we take a look at some AI techniques and try to demonstrate how google uses these.
In the end we will try to draw a conclusion about how innovative the project actually is. Further we'll try to ask what impact it already has today and what potential it could have for the future.

2. Google Lens

Google Lens is a technology developed by Google and was initially released in 2017. Through image recognition Google Lens tries to bring up relevant information of objects that the user scans through the smartphone camera.

The user can point the camera at any object and the Lens-App tries to develop an understanding of what the object is and then displays different information about the result to the user.
The Image on the right side shows how the basic UI looks when using the app.



It's use cases vary deeply. E.g. you can simply search for clothes you see and you want to know where to buy them. Or maybe in a foreign country you want to translate a food caption.
In one advertisement clip google demonstrates that the text-to-speech feature, which is integrated into google lens, can help people who struggle to read to understand the world around them.
The tool has many other features integrated which together build up to a great Project which might not yet be finished since not everything is working perfectly. But it shows how much potential AI has to simplify our lives and it can already be a big help in many situations.

3. Uso de técnicas de IA

Now I am going to discuss how Google made the functionalities of Google Lens possible and what concrete techniques were used in regards to AI.
Google wants the app to be able to operate on a wide range of devices. This is especially important, due to a big market in the developing countries and many illiterate people living there. For those people the project represents a solucion to be better integrated into everyday life (see introduction). To be able to deliver the performance needed to perform the requested tasks even on low-end devices, efficient algorithms are the key.
Now we are going to point out what algorithms Google Lens uses to implement its features.

Text-To-Speech (Artificial Neural Networks):
After the extraction of detected text Google Lens offers a text-to-speech (TTS) functionality which reads the text out loud. Of course this is not the only application of Google where this feature is used, regardless we want to take a closer look since it might be the most used functionality especially for people who can't read.

Generally there are two approaches to generate synthesized speech. Either by concatenating pieces of recorded speech (concatenative TTS) or by creating a completely "synthetic" output.
The latter approach has some important benefits, e.g. it is easier to modify the voice or express different emotions without recording a whole new database. Until a few years ago a technique called parametric TTS in combination with vocoders was used to generate this completely "synthetic" audio output. With this technique it is easy to alter the characteristics of the speech but in parametric TTS has tended to sound less natural than concatenative.

In 2016 DeepMind (Google) released a new technique, to generate pure "synthetic" output, called WaveNet. It brings major improvements in terms of intelligibility and naturalness of the voice. The technique makes it possible to change the speaker's identity, express emotions or imitate accents. Though in the beginning not efficient enough to be running on smartphones, it got many improvements over the last few years and now powers google's text-to-speech feature.
WaveNet is based on convolutional neural networks, a modification of deep neural networks which we will discuss in more detail in section four.

Other Ideas
fist part about how to process visual image and recognise entities (computer vision, R-CNN)
second part on how to generate the output (TTS, Knowledg Graph)

Optical Character Recognition:
Converts images of text into a machinex readable format.

Concrete (do only mention):
Regionbased Convolutional Neural Networks (R-CNN = RPN + CNN)
Region Proposal Networks detect regions of interest (ROI), a high-quality region proposal which tells the CNN where to look. To extract those regions a mechanism which is called selective search is applied.

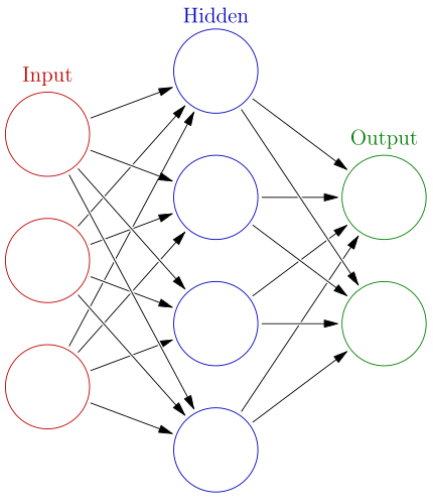
Image recognition
-> computer vision
-> Text recognition: optical character recognition (OCR), region proposal network (RPN), fast regionbased convolutional neural networks (Fast R-CNN)

4. Técnicas en profundidad

Artificial Neural Networks:

ANN's are a machine learning technique that is inspired by biological neural networks in brains. It's main components are nodes (called neurons) which simulate the biological neurons and edges which simulate the biological synapses.

Neurons hold a value which is calculated by the sum of all the values of the incoming edges and a specified function. Through edges to other neurons the value can be passed on to other neurons in the next layer (the neuron produces a single output, which can be passed on to several other neurons)
The Neurons of ANN's are distributed among three different kinds of layers.



An input layer, one or more hidden layers and an output layer, the signal usually flows from left to right. External input is received by the input layer. The output layer produces the ultimate result.

Mechanism
To receive a result for a given input by the network we first need to pass the input values to the first layer. Then respectively for every layer (one after another, from left to right) each neuron calculates its output value. First we calculate the activation of the neuron by calculating the sum of all inputs, weighted by a weight and adding a bias.
This activation value is then passed to an activation function which calculates the output. This is often a nonlinear function e.g. sigmoid σ (it maps to a value between zero and one). A little more explicitly the activation function could be defined as:

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

With x_1, x_2,... being the inputs, w_1, w_2,... the according weights and b the bias.

The inputs are the values of neurons with edges leading to the neuron at hand, or if the neuron is located in the first layer the input to the neuron is the input to the graph.
The weight describes the importance of the edge, to determine its specific value the network needs to undergo a learning process which we'll describe next in the learning paragraph.
The bias represents a constant added to the input, with which the value of the activation function can be shifted. Also the exact value of the bias is a result of the learning process.

As mentioned already, the output of the network is obtained by looking at the values of the activation function in the last layer. Usually we get a value for each neuron in the range of zero to one. If we are searching for one result, the node with the highest value represents it.

Learning
Now that we know about the general structure of the ANN, we can take a look at how such a network can learn to for example classify images.
First off we need a training data set (supervised learning), which will serve as the input to the network. During the learning process the network tries to adjust the weights and biases of all neurons and edges to minimize the observed error. This error represents the difference between the output of the network and the correct answer.

To calculate the error we need to define a cost function. This function will quantify how well the network is performing. A typical example would be the mean squared error (MSE), though this function is a crucial part when it comes to modifying the network to increase its performance.
The goal then is to find weights and biases which make the cost as small as possible. For this task the gradient descent algorithm or some variation of it is often used.

(Deep Learning)

Further Aspects

5. Innovación e Impacto

Google Lens itself doesn't invent any new techniques and neither is a revolutionary idea. Lens uses the current standard regarding many features like Image recognition (e.g. OCR), natural language processing (e.g. text to speech), or knowledge graphs (e.g. Google's knowledge graph).
That being said Google has created the leading technology for image recognition. Further Google has made multiple breakthroughs in the field of machine learning to make this product possible (e.g. Wavenet).
We therefore think that Lens represents as a product what Google's research teams have been able to achieve over the last years and that Google is a major pioneer when it comes to AI.

6. Bibliografía

https://d500.epimg.net/cincodias/imagenes/2018/12/05/lifestyle/1544009492_306342_1544009597_noticia_normal.jpg (ui image)
https://www.youtube.com/watch?v=ePwKgKp69GE (advertisement for google lens)
https://en.wikipedia.org/wiki/Region_Based_Convolutional_Neural_Networks (R-CNN, wiki)
https://deepmind.com/blog/article/wavenet-launches-google-assistant (wavenet, deepmind)
https://deepmind.com/blog/article/wavenet-generative-model-raw-audio (wavenet, deepmind)
https://deepmind.com/blog/article/high-fidelity-speech-synthesis-wavenet (wavenet, deepmind)
https://en.wikipedia.org/wiki/WaveNet (wavenet, wiki)
https://arxiv.org/pdf/1609.03499.pdf (wavenet, official paper)
https://analyticsindiamag.com/these-machine-learning-techniques-make-google-lens-a-success/ (lens techniques in general, aim)
https://analyticsindiamag.com/deep-learning-for-computer-vision-a-brief-history-and-key-trends/ (computer vision in lens, aim)
http://neuralnetworksanddeeplearning.com/chap1.html#eqtn6 (ANN's, Michael Nielsen)