Task 13

K Manoj
Referral Id:SIRSS2309

Text extraction from given images

Text recognition with machine learning

As you know, you need to teach the computer to recognize what we know is text. The task is a bit simpler when we talk about high-quality, legible pictures, where the text is clearly visible, and so are all the letters and digits. But what about pictures or scans of more mediocre quality? This is where the challenge begins. However, let's see how exactly does machine learning text recognition work.

OCR – Optical Character Recognition

First, we begin with the most common text recognition technique, and this is the OCR–Optical Character Recognition. OCR yields outstanding results only in very specific use cases, but in general, it is still considered as challenging. Optical Character Recognition is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera into editable and searchable data.

Let's say we have a piece of paper–a high school diploma. You can use your scanning device to put it into a computer, but it's not editable, for instance, with the MS Office tool. You need much more advanced graphics software to edit it. That takes time and requires specific skills. If you want to extract and repurpose data from this scanned document, you need an OCR software that would single out letters, put them into words, and then–words into sentences. This allows you to access and edit the document's contents at once.

The most advanced OCR systems are focused on replicating natural human recognition. The OCR systems are based on three main rules–integrity, purposefulness, and adaptability. First, the observed object has always to be considered as one entity comprising many interrelated parts. In our case, the

diploma is such an entity. Second, any interpretation of data must always serve some purpose. And finally, the OCR program has to be capable of self-learning.

The usage of the OCR software

The OCR software is by no means one, a uniform application that serves one and the same purpose. The OCR applications are used to serve lots of different intents. We can start with “reading” the printed page from a book or a random image with text (for instance, graffiti or advertisement), but we go on to reading street signs, car license plates, and even captchas. OCR software takes into consideration the following factors and attributes[1]:

- Text density. On a printed page, the text is dense. However, given an image of a street with a single street sign, the text is sparse. The OCR software has to recognize both.
- Text structure. Text on a page is usually structured, mostly in strict rows, while text in the wild may be scattered everywhere, in different rotations, shapes, fonts, and sizes.
- Font. While computer fonts are quite easy to recognize, handwriting font is much more inconsistent and, therefore, harder to read.
- Artifacts. There are almost none of them on a perfectly scanned page, but what about outdoor pictures? In short, this is a completely different story, and you have to keep that in mind when using OCR.

Real-world examples

Now, let’s consider two major examples for the real-world, outdoor conditions: House numbers and car license plates. House plates are extremely important, just to mention Google Street View and Google Maps. This is a massive source of tons of different house numbers. And as an example, Stanford University created out of them the SVHN (Street View House Numbers) dataset. SVHN incorporates over 600,000 digit images and is aimed at developing machine learning and object recognition algorithms.

Another widespread application of OCR is car license plate recognition. This also has a lot of possible applications, from police databases (data obtained from speed cameras) to private parking lots that open the barrier after a license plate is verified.

Machine learning text recognition in day-to-day situations

When a given car owner wants to leave the car park, they have to go to the ticket machine (or ATVM) and choose their number plate from the list. Right after the payment, the barrier management software receives a signal that the given car can leave the parking lot. When the car approaches the barrier, its license plate is scanned again, and if the scanned number matches the already-paid numbers list– the barrier opens.

This is an example of how machine learning text recognition can be extremely helpful in day-to-day situations. The car owner doesn't have to worry about a printed ticket and contrive where they should put it not to lose it. Everything happens within the software, and all the driver has to do is pay for their stopover.

The Future of ORC Technology

According to research published in April 2020 by Transparency Market Research, the global OCR market is predicted to be valued at $51, 527 million by 2030 and to expand at a CAGR of 15.2% from 2020 to 2030.

What to expect?

Today, engineers are working hard to discover innovative methods to integrate the essence of OCR into next-generation technologies. The answer is that the new generation of OCR is based on artificial intelligence (AI). This new form of machine-learning-led OCR can learn and analyze huge databases of extracting text from images, allowing the technology to think on its own. As a result, OCR technology is progressing from software that only scans and matches text to a program that identifies data and learns from it.

Text extraction from images using machine learning

With the text recognition part done, we can switch to text extraction. You see, at the end of the first stage, we still have an uneditable picture with text rather than the text itself. To solve this problem, the next step is based on extracting text from an image. Right after text recognition, the localization process is performed. All the related features about a particular image are gathered.

Text extraction: how does it work?

Text extraction, also known as keyword extraction, bases on machine learning to automatically scan text and extract relevant or basic words and phrases from unstructured data such as news articles, surveys, and customer support complaints.

The text extraction and enhancement methods are applied with the help of machine learning algorithms. And finally, the extracted text is collected from the image and transferred to the given application or a specific file type. There are many types of text extraction algorithms and techniques that are used for various purposes.

Therefore, we can divide them into five main methods

- REGION-BASED METHOD
  This method of text extraction uses a sliding window to detect text from any kind of image. This approach relies on several factors, such as color, edge, shape, contour, and geometry features.
- TEXTURE-BASED METHOD
  This method uses various kinds of texture and its properties to extract text from an image.
- HYBRID TECHNIQUE
  It's the combination of the previous two techniques. First, the region-based approach is used to detect a text. Then, with the usage of the texture-based method, all the features are extracted from the text region.
- EDGE BASED METHOD
  As its name indicates, this method is based on the detection of the edges of every letter and digit. This method is used to develop a high-level contrast between the text and the background.
- MORPHOLOGICAL BASED METHOD
  This method is used to extract all the text-related features from the processed image.

Use cases of text extraction from images

Every day, 2.5 quintillion bytes of data are generated by Internet users. A fascinating fact is that by 2020, each person generated 1.7 gigabytes in a single second. Comments on social media, product reviews, emails, blog articles, search queries, discussions, and so on. But the question is, how might text extraction from images help especially your company in becoming more efficient and take full advantage of the potential of data?

Social Media Monitoring

Your company can use text extraction from images to follow social media conversations to better understand customers, improve products, or take quick action to avoid a PR crisis. Text extraction from images may offer specific examples of what people on social media are saying about your business. Moreover, you may discover keywords and track trends with text extraction from an image.

Business Intelligence and Text Extraction from Images

Text extraction from images can also be effective in business intelligence (BI) applications such as market research and competition analysis. You may also get information from a variety of sources, including product reviews and social media, and participate in discussions on topics of interest. Furthermore, you can compare your product reviews with those of your competitors using text extraction from images and other text analysis tools. This helps in getting information that will help you in making data-driven decisions to improve your product or service.

Conclusion

To sum up, there is increasing demand for text extraction from images now. Many extraction techniques for retrieving relevant information have been developed. So, to successfully use text extraction from an image in your business, you should identify business goals, analyze data accessible from both open source and private datasets. Additionally, you should decide whether extra security measures are required to confirm a failure in the accuracy of the OCR mechanism.

Do you think that text extraction from images using machine learning might be beneficial to your company or speed your work up? Don’t hesitate to call us or send us an e-mail. Addepto is a professional Machine Learning Consulting company. We are very keen to talk with you about implementing text extraction from image solutions to your business.