Designing of a scene text recognition system

Technology, Artificial Intelligence

The current wave of industry digitalization is known as Industry 4. 0, involving cognitive technologies such as robotics, artificial intelligence, sophisticated sensors, chatbots, cloud computing, internet of things (IoT), data analytics, digital fabrication, cyber-physical systems, communications, mobile devices, and others. These technologies are jointly utilized to integrate the physical and virtual environments. The paradigm of industry 4. 0 is intended to construct the factories of the future – so called smart factories. Over the industrial internet of things (IoT for industry), various cyber-physical systems monitor and control the physical processes, communicate and cooperate with each other and with humans in real time. Automation and robotics are the basis of many trends including machine vision systems for robust identification processes and a high detection rate. Many industries utilize machine vision systems to provide automatic image-based inspection and applications analysis such as automated inspection, process control and robot orientation. The combination of machine learning techniques helps improve the machine vision systems, to identify and recognize patterns based on computer vision techniques. The object recognition and the text recognition are a sub-domains of computer vision techniques. Deep learning is the most successful technic for many applications in the computer vision domain. Convolution Neural Networks (CNNs) have improved results on object recognition and object detection and enabled industrial applications.

Heavily Adoption of deep learning and Artificial intelligence by cloud providers and enterprise companies leads to provide their customers with value added services. So, a serverless computing environment is suitable for

the inferencing of large neural network models. Scene text recognition has attracted great amount of research interest of the computer vision community due to its numerous applications. Scene text detection in natural images is a crucial step towards end-to-end scene text recognition. Identification and extraction of scene text from natural images and videos have been an essential job in recent computer vision research due to the frequent usage and advent of smart gadgets. There are several primary reasons for this trend, including the demand for a growing number of applications. Text is one of the most expressive means of communications and can be embedded into documents or into scenes as a means of communicating information. This is done in the way that makes it “ noticeable” and/or readable by others. For e. g. The collection of massive amounts of “ street view” data is just one driving application. The second factor is the increased availability of high performance mobile devices, with both imaging and computational capability. This creates an opportunity for image acquisition and processing anytime, anywhere, making it convenient to recognize text in various environments. The third is the advance in computer vision and pattern recognition technologies, making it more feasible to address challenging problems. Due to the frequent use and advent of smart gadgets, the identification and extraction of scene text from natural images and videos has become a regular job in recent computer vision research. It also has a huge demand in content-based image retrieval and understanding.

Technically, text extraction undergoes through two major steps: (a) Text Detection in which it is identified and localized from natural scenes and/or

videos. In summary, it’s a process to determine text/non-text regions. (b) Text Recognition means to understand semantic meaning of the text. In general, two types of text exist; (1) scene text which is normally a click of camera and reveals common surroundings. This makes scene unstructured and ambiguous due to uncertain situations e. g. , advertisement holdings, sign boards, shops, text on buses, face panels and many more, (2) Caption or graphic text is added manually to images and/or videos to support visual and audio content, making it a simpler text extraction method compared to natural scene text.

Text recognition in the natural image is still a challenging task due to complicated environment. The challenges regarding complexity of natural images for text extraction can be broadly seen from three different angles.

1. There may be variation in natural scene text due to uncontrolled surrounding which reflects absolutely different font size, style, color, scales and orientation,
2. Scene background complexity have challenges like roads, signs, grass, building, bricks and paves etc.
3. Some intrusion factors like noise, low quality, distortion, non-consistent light also creats problem in natural scene identification and extraction.

Problem such as blur, text features weakened or lost exist in text region because of dust on images. As the Text recognition gives rise to numerous applications, the fundamental aim is to determine whether or not there is text in a given image and if it is there then the problem is in detecting,

localizing and recognizing it. So, for text detection, pre-processing and post - processing is mandatory task.

Mostly Text enhancement is used to rectify distorted text or improve resolution. The analysis of challenges of text detection in given image is can be given as: Scene entanglement: The challenge with scene complexity is that the Environmental scene makes it difficult to discriminate text from non-text. Many man-made objects, in natural environments, such as painting, symbols and buildings, appear that have similar structures and appearances to text. Example like character ‘ Z’ can be seen as a design on gate of houses, character ‘ O’ can also be seen in given house image. Improper lighting: Sensory device’s uneven response and illumination is the main cause of improper lighting at the time of capturing images. Because of uneven lighting colour distortion and deterioration of visual features false detection, segmentation and recognition results. Blurring and degradation: Flexible condition of work, focus less cameras, image compression - decompression procedures defocuses, blurs and degrades the quality of text. These factors reduces sharpness of characters and increases number of touching characters, making basic tasks difficult.

Aspect ratios: A text can have different ratios of aspect. Text like signs of traffic, may be shorter, but other text, like video captions, may be much longer.

Distortion: Perspective distortion occurs if camera’s optical axis is not perpendicular to the plane. Because of this, bounding boxes of text lose the

shape of rectangle and character also gets distorted. So, the performance of recognition models trained on undistorted samples decreases.

Fonts: Characters of some particular language’s fonts may overlap each other, so segmentation task will be difficult. Characters of various fonts have larger variations and form many pattern sub-spaces. So it becomes difficult to get accurate recognition whenever the character class number is high.

Multilingual environments: Every language have various characters. Some languages like Korean, Japanese and Chinese, have thousands of character classes. Some language like Arabic has connected characters, where shape can be changed according to context. Language like Hindi combines alphabetic letters into thousands of shapes that represent syllables. Because of the challenges of text detection dataset have to be generated in huge number. Then only output will be more accurate. For this most popular technic deep learning model can be used. But model training has became more time- consuming process because of availability of large dataset and increasing complexity of deep learning model. In many application like Computer vision, Natural language processing, and Speech recognition Deep neural network is the most successful technic. Specifically, in the computer vision domain, Convolutional Neural Networks have improved results on object detection, recognition and enabled industrial applications. For this, Training need to be done over larger dataset because model has million parameters and so complexity is also increasing day by day.

Training a CNN model is a time-consuming process. For speeding up this process three criteria should be considered. First, specialized processors,

(Graphics Processing Units (GPUs), TPUs etc.) and software libraries (CuDNN, fbfft) can be used. Many of the popular open source Deep Learning (DL) frameworks now offer distributed versions that allow the user to train models that utilize multiple GPUs and even multiple nodes. Some comparison which can be shown for different distributed versions stack up against each other. We can also measure the quantitative and qualitative performance in terms of time, memory usage, and accuracy for Caffe2, Chainer, CNTK, MXNet, and Tensorflow as they scale across multiple GPUs and multiple nodes. Among various frameworks of deep learning, open- source packages that support distributed model training and development, gives more proper results.

Five selected framework are given as follows:

1. Caffe2 is a light-weight and modular DL framework open sourced by Facebook. It emphasizes model deployment for edge devices and model training at scale.
2. Chainer is a flexible DL framework developed by Preferred Networks that provides an intuitive interface and high performance implementation. The distributed version is ChainerMN. Rather than separate the definition of a computational graph from its use, Chainer uses a strategy called “ Defined-by-Run” where the network is created when the forward computation takes place.
3. Microsoft Cognitive Toolkit (CNTK) is a commercial grade distributed deep learning toolkit developed at Microsoft. It also has advanced algorithms but these are not under open-source licenses.

4. MXNet is a flexible and efficient library for deep learning, featuring high-level APIs. It is sponsored by Apache Incubator and selected by Amazon as its choice for DL.
5. Tensorflow4 is a general numerical computation library for data flow graphs. It was developed by Google Brain Team and is currently an open source project for machine learning and deep learning domains. For the implementation of a big range of tasks based cloud applications, serverless computing is best.

So, finally this system will get integrated with AWS plateform using Deeplens. Amazon Web Services is a part of Amazon. com that gives according to demand, cloud computing platforms. It includes on-demand delivery of compute power, applications, database storage, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing.