White Paper

Achieving a Better Understanding of Object Detection Using Deep Learning Methods: Opportunities and Limitations for Media Applications

Sanjay Rao
Interra Systems, Inc.

1. INTRODUCTION

Object detection is a method of locating and identifying when certain types of objects such as computers, animals, and vehicles, appear in an image using computer technology. Usually, these methods fall into one of two categories: traditional machine learning (ML) or deep learning.

ML-based methods require expert human involvement, in terms of defining image features and then using techniques like support vector machine (SVM) to classify the image based on the features.

Deep learning-based methods do not require feature definition. Given enough data, these methods are capable of automatically learning the relevant features for the task at hand. These methods can perform end-to-end object detection. They are typically based on convolution neural networks (CNN). Deep learning-based methods use multiple layers of convolution neural networks, which are trained using a dataset of the objects of interest. Based on the dataset used during training, the network learns different features of the image and uses these features to locate and identify the objects.

The hype created in the media about AI and deep learning has led to a scenario where many people believe we already live in the future and everything has been infiltrated by AI and ML. It is undeniable that deep learning has been quite successful in computer vision tasks like image classification and object detection. But the idea that given enough data, ML can solve all the problems is false. This belief actually harms the value of machine intelligence by setting up unrealistic expectations. Instead of painting an unrealistic picture of ML, it is important to set the right expectations so that people can make informed decisions. Integrating ML solutions with an accurate set of expectations will result in a better outcome.

This paper aims to explain some of the basic limitations of deep learning-based methods and corrective measures that might help improve the outcome of any ML-based solution. The problem of object detection has been used as an example wherever required.

At a granular level, accuracy of any classification algorithm is measured using three parameters called recall, precision and F1 score. Recall represents the fraction of true positives among the total relevant instances. Precision depicts the fraction of true positives among the total retrieved instances by the algorithm. F1 score is harmonic mean of precision and recall. Possible values of these parameters can go up to one. Figure 1 explains these parameters visually.

The accuracy of an object detector is slightly more complicated and measured using mAP (mean average precision). mAP includes accuracy with respect to both location and classification of the object.

In publications, different algorithms and research papers publish accuracy numbers on specific datasets. These specific datasets are generally simpler for classification. A popular algorithm that gives 0.9+ F1 score on a specific image dataset, comes down to 0.25 for the video dataset on movies and advertisements.

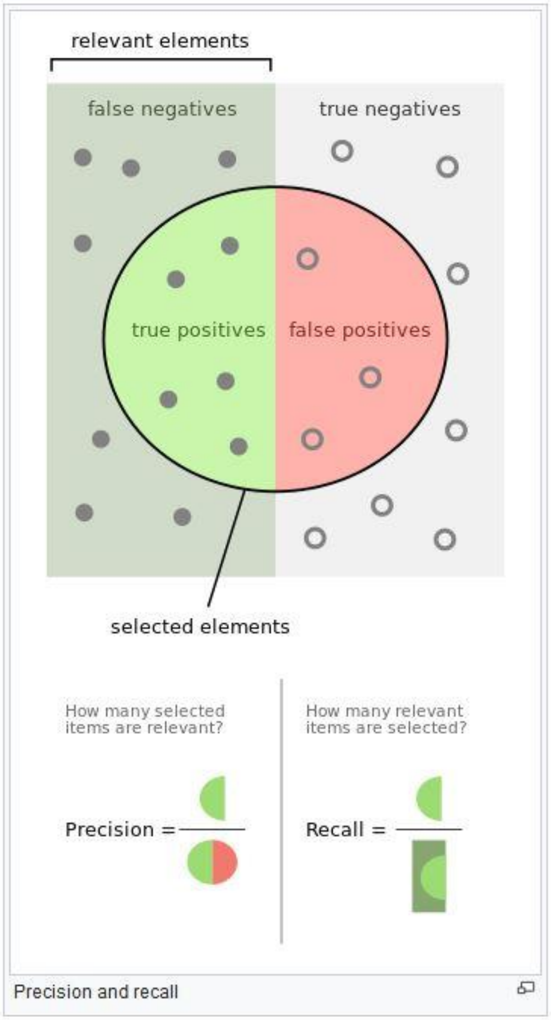

Figure 1. Graphic representation of precision and recall.

2. LIMITATIONS OF MACHINE LEARNING (OBJECT DETECTION)

Judgment

All deep learning-based algorithms are developed using copious amounts of data. In the case of object detection, these algorithms are trained using a number of images having an object of interest. During training, deep learning algorithms learn to extract a useful feature set from the images. In a good object detection algorithm, this feature set is large enough to distinguish the object of interest from other objects present in the training images. However, the object detector learns to identify patterns in images using training data without developing any real judgmental capability. Such an object detector can generate interesting results in certain cases.

In contrast to this, human interpretations follow rules that go beyond technical prowess. Before looking into any image, humans have extensive background knowledge about the scene in the image, objects present in the image, and more, through past experience. This background knowledge helps humans identify the object correctly, even when an object of interest is not fully visible or its individual attributes are not directly discernable from the image itself. For example, in Diagram 1 below, a human being is likely to judge that the child is not holding an alcoholic drink. While inferring this, a human will use background knowledge about the other aspects like:

- In this simple situation, this young boy cannot have an alcoholic drink.
- By looking at the color and pattern of the can, along with the partial logo, a human being can determine that it's a soft drink brand.

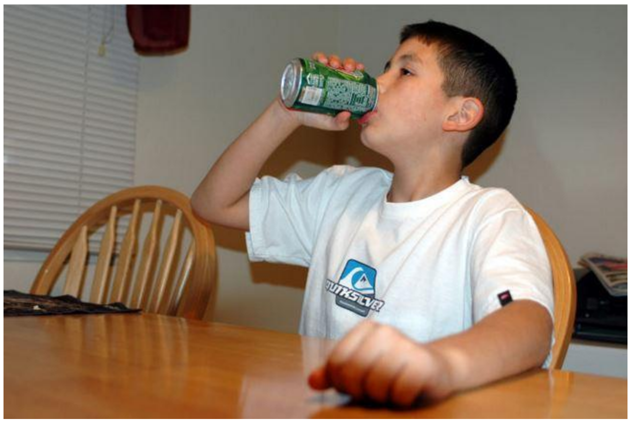

Diagram 1. Example of the impact of human judgment on object detection. Courtesy: www.travis.af.mil

Often there are environmental aspects and facts that might not be present (or accounted for) in the training dataset. This will affect the object detector's output. Though researchers are trying to incorporate these aspects in the object detector techniques, the number of aspects is practically limitless.

Developing human level understanding in an object detector is very difficult, if not impossible.

Interpretability

The results of deep learning techniques are not interpretable. With current approaches of deep learning methods, it is not possible to decipher how the algorithm came to the decision it did. Most of the approaches in object detection and other ML techniques suffer from interpretability issues, despite their apparent success. In the absence of interpretability, it is not easy to convince business clients and investors that the method is accurate and reliable.

Data

Deep neural networks are data-devouring beasts and require copious amounts of data to train an object detector. Big companies like Google and Facebook have access to this amount of data and large teams to annotate them, but it is not easy to acquire this amount of data in all cases. Augmentation of data is one approach to solve the problem, but augmentation has its own limitations.

Lack of Good Data

Even if you have a lot of data, covering all of the variations (e.g., size, color, background, orientations) of the objects is nearly impossible. Each good object detector requires a large amount of handcrafted, structured training data.

Data-Driven Approach

Theoretically, any ML algorithm can achieve accurate results after training on an infinite amount of data. If a model has not been trained on a particular type of data, it cannot reliably predict the output for that type of data. Sometimes a very minor change in the image completely alters the results of the object detector on the image. Diagram 2 shows a popular image classification model’s output, before and after adding some noise to the image.

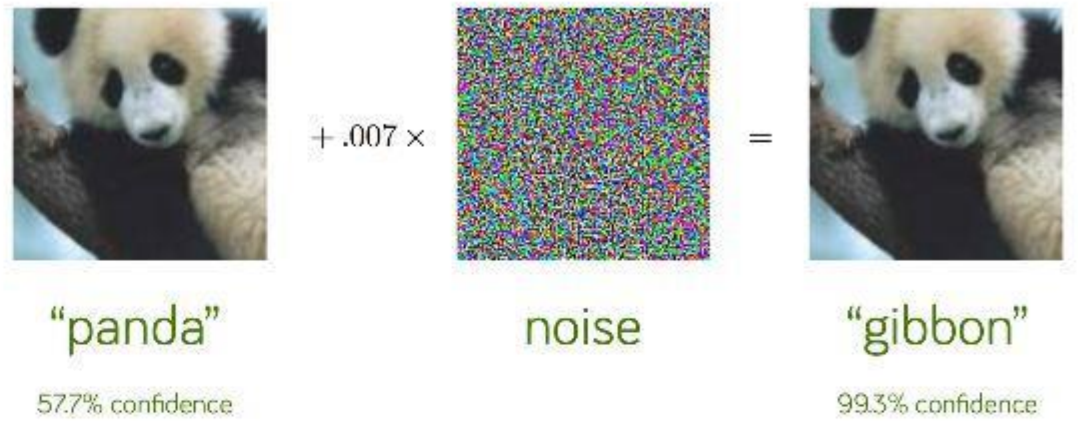

Figure 2. Image output before and after adding noise to the image. Courtesy: Explaining and Harnessing Adversarial Examples

Another simple case is if an object detector has not been trained using color blended images then it cannot confidently identify the clearly visible objects in the blended images. (For an example check out YouTube). The same thing is possible in multiple situations, and some examples are shown in Diagram 3. Fundamentally, ML provides computers with the ability to learn without being explicitly programmed, but this learning is static and is difficult to transfer well in new situations.

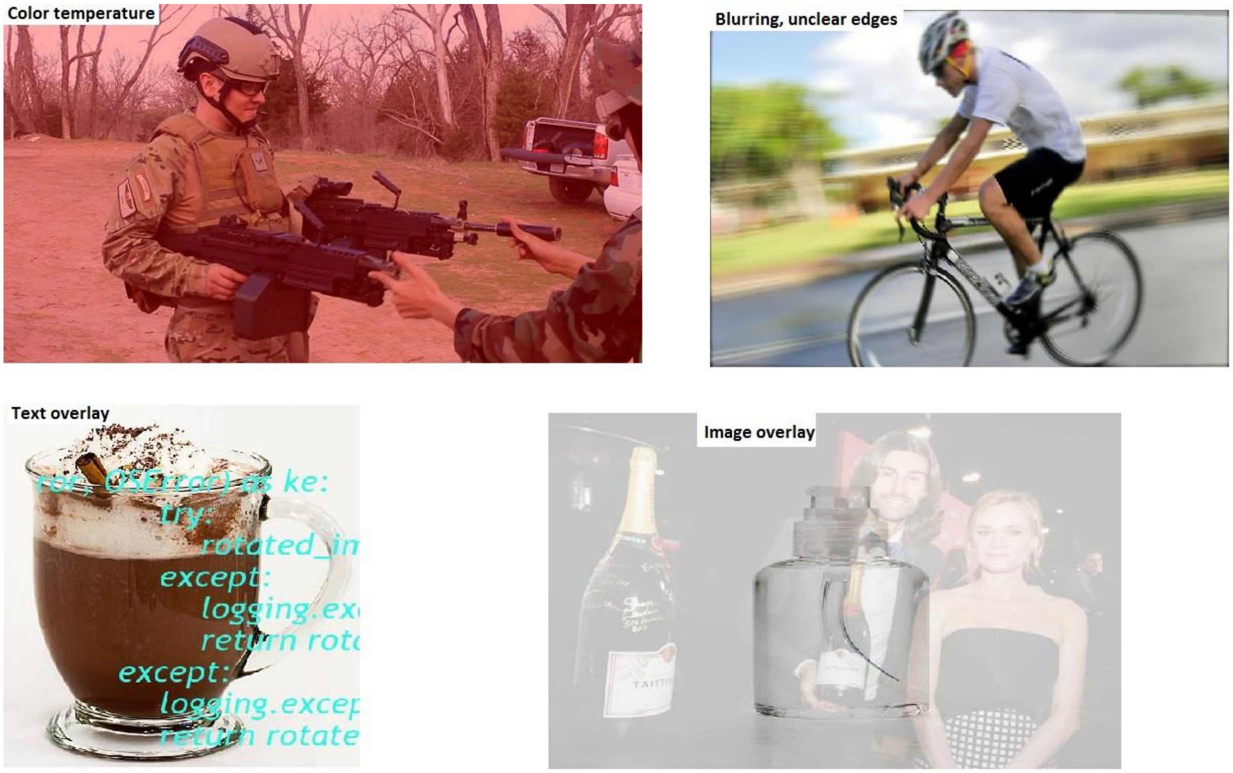

Diagram 3. Examples of blended color images

Stochastic Approach

Deep learning is a speculative (i.e., stochastic) approach, not deterministic. The model uses the inputs available to it. Based on the provided inputs, it gives a speculative result. So by definition, it is not meant to be 100% perfect, because in any moderate level problem data might not be available for all possible scenarios or data could be conflicting.

ML approaches encode correlations between different features of data using the training data, trying to maximize the correctness of the output on training data. These approaches do not encode causation and ontological relationship.

Deep learning-based techniques are speculative approaches to identify whether an object is present or not. More optimistic speculations result in more false positives, and less optimistic speculations result in more false negatives.

An object detector trained using abundant data can learn spurious correlations between thousands of features of the images. Based on these correlations, the object detector can behave well on training data but badly on test data. This is similar to p-hacking phenomena, where researchers select or collect data until a desired non-significant result becomes significant on the data.

Other Common Problems

Illumination and Background

A deep learning network cannot generalize itself easily to the variations in illumination and background. In other words, if the training data contains an object only in a specific background (e.g., dogs in outdoor conditions), then it may not be able to detect the same object in other backgrounds (e.g., dogs in indoor conditions). Training data needs to cover a wide variety of background and illumination.

A complex background generates numerous patterns in the image. One or more patterns could match one of the interesting objects, resulting in false positives.

Appearance and Object Similarity

Objects are 3D whereas images are 2D. A 2D projection of an object could be very similar to another 2D projection of a different object. It leads to a false-negative for the first object and simultaneously generates a false positive for the second object. There are some objects that are highly similar in shape and properties. For example, alcohol bottles and olive oil bottles.

Occlusion and Blending

Partially occluded objects cannot be identified correctly without involving the possibility of false positives. A blending of two images creates effects that object detection is unable to negate. The object detection method cannot distinguish into two separate images and could result in either a false positive or a false negative.

Image Quality Issues

Blurriness and especially motion blurring in videos distorts the features of the image. The resultant value of features deviates from the values on which the network is trained. This again can result in false positives or false negatives. There could be hundreds of other image quality problems to cover, a mountainous task for an object detector.

3. CORRECTIVE MEASURES

Irrespective of all these limitations, the ML field has good potential to solve object detection and many other computer vision-related problems. These insights of ML help us understand what is possible using ML and to what extent. There are several corrective measures that can be used to reduce the impact of the limitations and improve the outcome of ML techniques. A couple are discussed below.

Manual Reviews

Deep learning's current role should be more of an implicit partnership with humans. With all the current limitations, it still leaves room for humans to think smarter and innovate more.

In a ML-based solution, wherever possible, there should be a variety of options to quickly review the results generated by deep learning networks. For a long time, popular social media platform companies believed that the spread of misinformation and hate speech could be algorithmically identifiable. Under pressure from legislators, they hired a large number of human content moderators and started using the algorithms as assistance to speed up the job.

Some problems arose in the legal domain and in the U.S. when algorithms were used to sentence criminals. The algorithm calculated risk score and advised judges on sentencing. The algorithm was found to amplify structural racial discrimination and was later abandoned.

Avoid Online Training

Do not provide critical solutions to learn incrementally or interactively. Offline/batch training using handcrafted data generally performs better. With all of the above limitations in ML, online learning could be very risky in production applications.

4. CONCLUSION

It is undeniable that deep learning has opened up a plethora of opportunities for solving different problems in various fields. Within certain markets, in particular the video and entertainment industry, ML-based object detection can help increase accuracy and efficiency. For example, using ML-based systems, media companies can quickly generate audio-visual descriptors from video files, including identification of faces, people, violence, logos, various objects (e.g., firearms, cold weapons, cigarettes, body parts, and alcohol), visual text, language, spoken dialog, nudity, and explicit scenes, with a high degree of confidence.

However, it has also led to an emergence of a mindset that ML can solve many complex problems without much human intervention. The aim of this paper is to make it clear that there are limitations that prevent that from being the case, at least for the time being. A good understanding of the inherent limitations of the ML methods helps us devise ways to use the technology more effectively.