GE
Intelligent Platforms

Classification of objects from video streams

Classification of objects from video streams

Functional need

The human brain remains the most effective classifier on earth, able to perform real time differentiation between a huge number of disparate objects and events, in varying environmental conditions of lighting and occlusion. This ability is learned from birth, and while there is significant overlap of experience leading to the illusion that we all classify in the same way, each human is in fact operating in a highly individualistic manner.

In fact, the classification is actually part of a wider cognitive process in humans and so outlier objects, behaviors and events can be interpreted differently by different individuals depending on their prior experiences. Furthermore, this interpretation is shown to be affected by many factors including the emotional state of the observer, and sensory loading. This is most notable in the evidence given by multiple witnesses to events that fall outside the "normal" human experience, such as major accidents or crimes; in that moment, each human cognitive process has been highly focussed on processing the new and unusual events, meaning that normally reliable functions such as memory or behavioral reactions have become secondary, leading to questionable subjective recollection.

Efforts to replicate the human ability to classify have historically been largely academic, lab-bound and isolated; each employs one chosen algorithm, is often coded, tuned and operated on the same platform; and does not apply continuous learning or evolution.

In the world of the Industrial Internet, there is an opportunity to have many different machines observing the physical world through a variety of sensors and applying a "collective learning", such that the experience of the individual machine contributes to the learning of the whole. In this model, the processing capability exists in large data centers at the heart of the network; here, there are sufficient computing resources to tackle the problem of algorithm tuning and continuous learning.

The edge nodes of the network interact with the real world sensors and apply processing capability to the problem at hand, classifying, filtering and reducing the vast quantity of sensor data into a useful set of information to pass to the rest of the network, as well as identifying errors and failures in the algorithms. The datasets associated with these failures can be transmitted to the central resource for analysis and further learning to improve the algorithms (and hence performance) at the edge nodes. The benefit of the model outlined above is that the learning experience is not confined to a single machine, but is distributed for the whole network to benefit from.

Current state of the art

There have been many years of research into video and image processing, with a great many mathematical strategies proposed to solve a variety of problems. Early research was hindered by the lack of computational resources, requiring huge and expensive machines to run the coded algorithms on stored video clips and image stills. As Moore's law predicted, the ever-increasing density of integrated components means that processing performance gets greater each year; the capability of ICs designed for the mobile industry now puts colossal computational capability into a size, weight and power (SWaP) envelope that ensures video and image processing will soon be ubiquitous in a wide variety of applications.

One example of such an application is the work undertaken by NVIDIA to implement advanced driver assistance systems (ADAS). Here, the edge nodes of the system

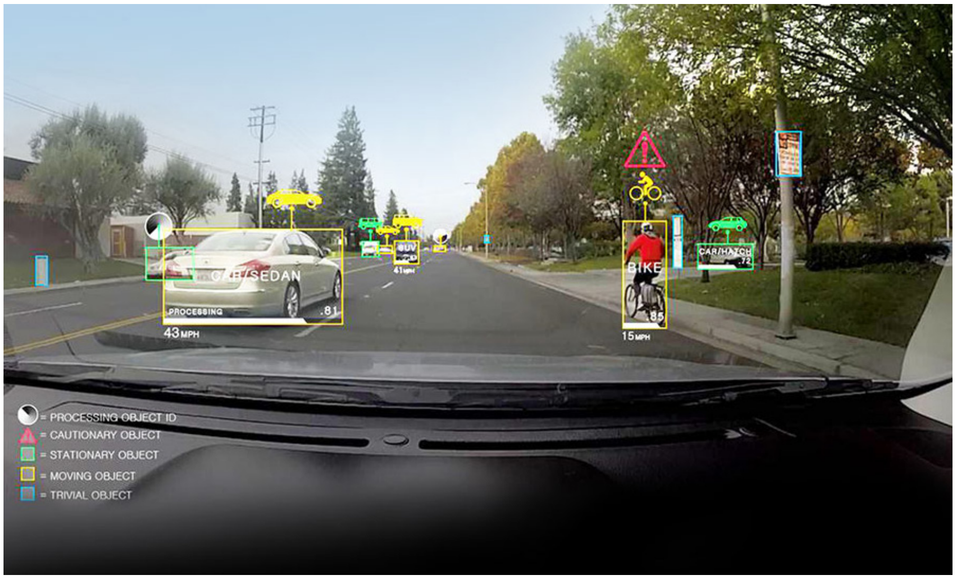

Figure 1 – Object classification using NVIDIA's Advanced Driver Assistance System (Source: NVIDIA)

are highly-performant System-on-Chip (SoC) processors mounted inside individual cars. The SoC runs several different types of algorithms to classify objects and also determine something about their motion; pedestrian versus car, stationary car versus moving car. The objects and situations that cause the classifier to fail will be uploaded to the cloud for analysis on huge clusters, and the resulting improved classifiers sent to all cars in the network. The result is that every car learns simultaneously from the experience of an individual. The network is objective in its response to a particular object or event, regardless of the situation. (Figure 1)

The latest NVIDIA SoC, the Tegra X1, is a "mobile supercomputer" comprising a quad-core ARM architecture coupled with 256 graphics processing cores based on the "Maxwell" architecture. This combination makes the processor extremely effective at handling vision-based tasks based on parallel processing of the large datasets produced by video cameras. The NVIDIA Tegra X1 is rated at 1 TeraFLOPS of processing performance, with a single computer module measuring the size of a credit card and consuming less than 10 Watts implementing this SoC. This is remarkable, given the fact that the first computer to reach 1 TeraFLOPS of processing performance was the ASCI Red supercomputer built by Intel and Sandia National Laboratories in 1996; it measured 1,600 sq ft and consumed 850 kiloWatts. ASCI Red remained the most powerful computer in the world until 2000. In the not too distant future, every car on our roads is likely to carry several TeraFLOPS of processing performance!

Classification Algorithms

There are many image classification algorithms. Some of the common algorithms are discussed below.

Support Vector Machine (SVM) is a "supervised learning" model, in which the system is given training examples that fall into a category and others that do not. The algorithm maps the examples into a multi-dimensional space, and then attempts to find a plane through that space with the greatest separation between those in the category and those not. New examples are then mapped into the space and classified according to which side of the line they fall.

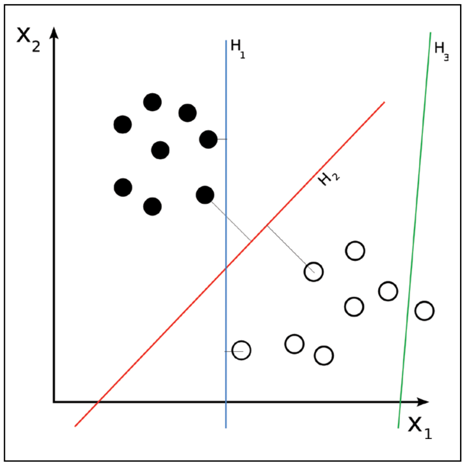

Figure 2 – SVM classification. H1 has a limited distance to the two groups; H2 achieves maximum distance to the two groups; H3 does not separate the two groups

The SVM algorithm dates back to the 1960s, but research has led to progressive improvements in it, including dealing with misclassification and non-linear solutions. However, the quality of the solution depends on the quality of the initial example dataset. (Figure 2)

An alternative to SVM is the Cascaded Haar algorithm, which is particularly suitable for classification of objects within an image. The system is again trained using a set of example images. A series of classifiers detecting edges, lines and center-surround features are scanned across the image, returning a positive or negative result if the feature matches the image. The algorithm applies multiple classifiers to the image until an object is either recognized or rejected. The classifiers may be grouped into complex classifiers via a number of weighting techniques. This cascade of classifiers is then applied to the candidate image until all the classifiers are either passed or rejected. (Figure 3)

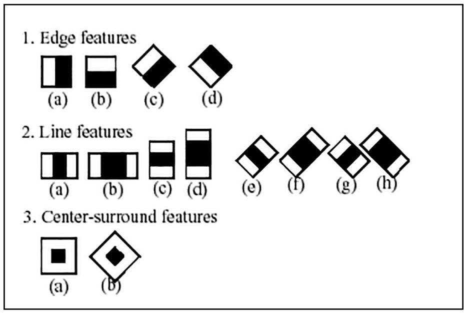

Figure 3 – Haar-like feature classifiers

There is currently much implementation effort being applied to machine learning using Deep Learning techniques; sophisticated, multi-level "deep" neural networks (DNN) optimized for GPU. This technique uses nodes, or "neurons", connected together to mimic the biological neural network found in the human brain.

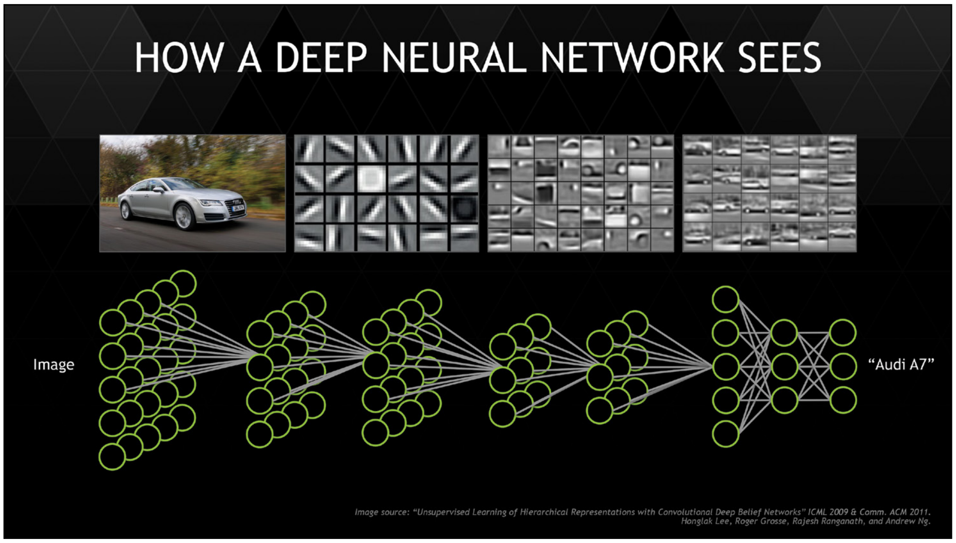

Image source: "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks" ICML 2009 & Comm. ACM 2011. Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng.

Figure 4 – Schematic representation of a deep neural network, showing how more complex features are captured in deeper layers. (Source: NVIDIA)

During learning, the connections develop adaptive weightings to simulate the strength of connections between neurons. Layers of neurons connected together translate their input via the weighting to the next layer. The more complex the system, the more layers that need to be implemented and the more processing performance that is required; hence, various techniques exist for optimizing the connectivity. DNN algorithms are very well suited to object classification for images and video. (Figure 4)

The neural network may be trained to recognize many objects under different conditions, but due to the way in which the weightings must propagate both forwards and backwards through the network, the learning phase is very computationally intensive. In deployment, the network only propagates in the forward direction, and so can be run much more efficiently on small, deployable hardware.

In addition to these object classification techniques, inter-frame temporal processing allows for added information about the objects to be derived; are they static within the scene, or are they moving, and with what velocity relative to the viewer?

Implementation

The implementation of object classifiers on deployable hardware relies on high performance hardware and very well optimized libraries. GE Intelligent Platforms has selected the NVIDIA Tegra System on Chip as the basis for a new product because it meets both of these criteria. The first rugged product, the mCOM10-K1, is a credit card-sized computer-on-module using the Tegra K1, with quad-core ARM A15 CPU and 192 "Kepler"-architecture GPU cores, delivering 327 GFLOPs in a 10 Watt power envelope. Future developments will incorporate the recently-announced Tegra X1, driving up the performance and increasing the functional capability.

Meanwhile, NVIDIA is continuing to invest in the development of highly optimized libraries to support object classification, including the OpenCV Open Source Computer Vision library, the cuDNN CUDA Deep Neural Network library, and the underlying CUDA itself.

GE draws on both the SoC and the optimized libraries to create compelling products and demonstrations of this technology, and a roadmap to ensure that investments in technology today are protected into the future. (Figure 5)

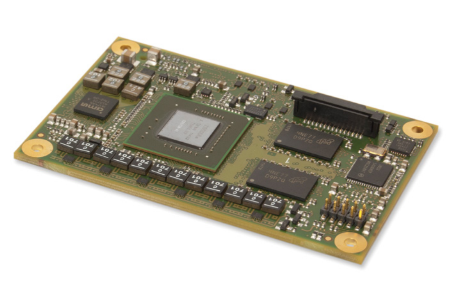

Figure 5 – Measuring 84x55mm, GE's TEGRA K1 computer on module

Future concepts

Along with an ongoing technology insertion roadmap, GE is integrating these computer vision functions with other key technologies into a broad range of advanced applications that will be transformative to our lives. For example, a computer-based classification system linked to a Data Distribution Service (DDS) or OLE for Process Control (OPC-UA) networking middleware will allow for real time exchange of data between different edge nodes within a larger community of computers. The individual pixels are not needed in the larger system network in order for the information about the detected objects to be disseminated and correlated, even when observed from different viewpoints. Assuming video can be stored on individual vehicles, it would also be possible to interrogate each edge node for retrieval of video clips recorded at a geo-specific location before and after a specific trigger, allowing for a comprehensive review of events leading up to and following an accident.

As learning algorithms break free from the confines of the computer rooms of research institutes and into the real world, the evolution of these algorithms is expected to progress rapidly, allowing classifiers to distinguish between an ambulance and a delivery truck, or a school bus and a van, for example. The roadmap of low SWaP computers with colossal processing performance is set to keep pace with the demands from the researchers' algorithms and propel us beyond object classification into complex behavioral classification. The ADAS systems of the future will all objectively understand that a small human is more likely to run into the road than a large human, and that probability is increased if the small human also has a ball...

Imagination at work

03.15 GFT-887