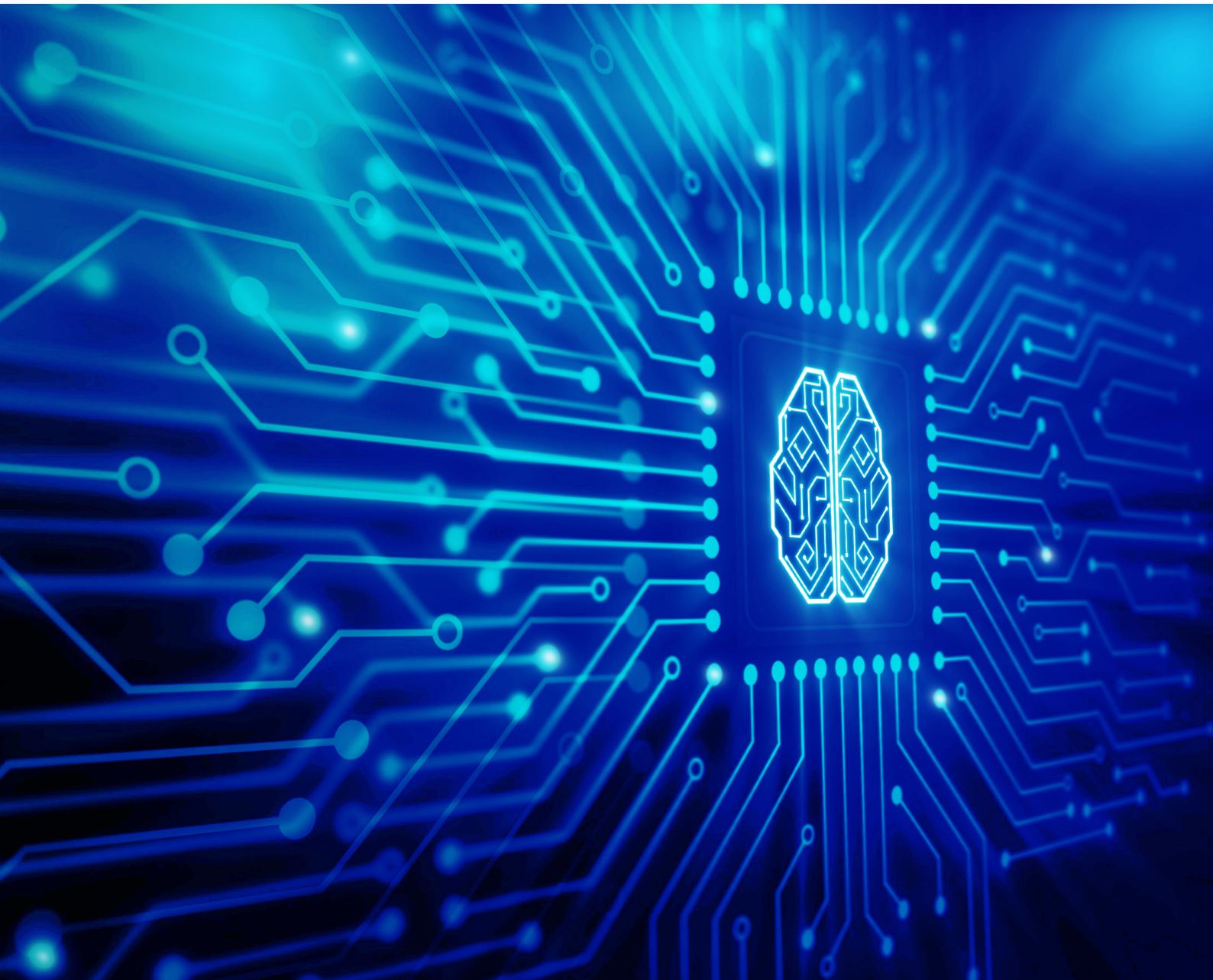
Calipsa eBook

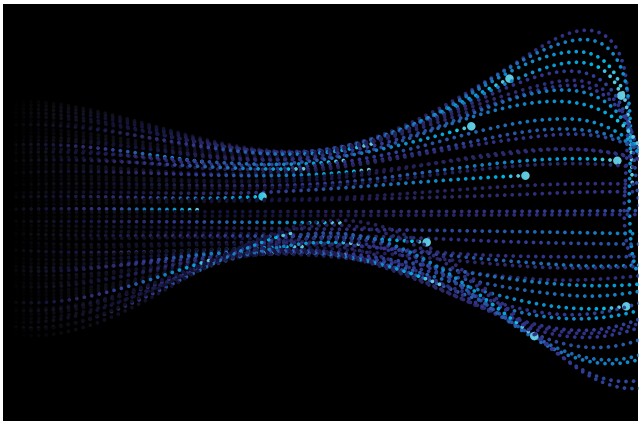
MACHINE LEARNING EXPLAINED

The 3 essentials for video analytics



CONTENTS

INTRODUCTION



Whether you're familiar with Calipsa's False Alarm Filtering Platform or not, it's likely you've heard a lot about machine learning and artificial intelligence over the past few years.

At Calipsa, we use machine learning to help monitoring companies make sites safer and prevent crime from taking place. But how exactly does machine learning come into this - and how does it work?

In this ebook, we explain the basics of machine learning, breaking down the essential elements that every system needs to get started. Along the way, we'll explain how we use these elements at Calipsa to make our video analytics platform so accurate at improving crime prevention.

WHAT IS MACHINE LEARNING?

Before we dive into all things machine learning, let's start with the basics. **Artificial intelligence (AI)** is any technique that enables a computer to mimic human intelligence; some of these techniques include logic, if-then rules, decision trees, and machine learning.

Machine learning is just one type of artificial intelligence. Machine learning algorithms use statistics to find patterns in vast amounts of data, and then they apply those patterns to make predictions. The more data the algorithms work with, the better they get at spotting patterns, so they "learn" to improve at tasks with experience.

The recommendations you get on streaming services like Spotify or Netflix are real-world applications of machine learning. For example, if you watch Planet Earth on Netflix, you are telling the algorithm to look for similar shows to recommend.

As well as looking at the viewing patterns of other people who watched Planet Earth, the algorithm will also draw on information such as the genre, presenter, and year it was made, to suggest new shows for you to watch.

Deep learning is a subset of machine learning where the algorithm trains itself to perform a task by exposing deep neural networks to vast amounts of data. That's a lot to take in, so let's break it down.

Deep learning depends on neural networks, so-called because they were inspired by how biological neurons work in our brains. The human brain can solve a complex problem, but individual neurons are each responsible for solving specific, smaller parts of the problem as a whole.

Similarly, a neural network is made up of many layers of computational nodes. Each node is responsible for solving a small part of the overall task at hand, a bit like a neuron in the brain. This gives computers an enhanced ability to spot patterns, as all the nodes work together to process much larger amounts of complex information, and make a prediction.

WHY IS MACHINE LEARNING SO IMPORTANT FOR VIDEO ANALYTICS?

When we watch a video, we take for granted that highly sophisticated processes are taking place between our eyes and brain in fractions of a second, without any effort from us. Something as simple as watching someone sit in a chair involves multiple levels of understanding: what a person looks like, what a chair is, that the downward motion we're watching is the transition from standing to sitting.

So how do we get a computer to do this - and why would we want to?

Let's start with why. The security industry has traditionally been heavily dependent on manpower, even where video technology is concerned. Until now, guards and control room operators have either had to monitor hours of continuous video footage, or review the alarms that security cameras send through. In both cases, people could be using their time and expertise much more effectively.

Guards and operators are on the lookout for **genuine alarms** - this is footage that contains human activity in a scene where people aren't supposed to be. However, around 95% of the alarms security cameras send are **nuisance alarms**; in other words, these alarms don't contain any human activity, but operators still receive them. Nuisance alarms can be triggered by all kinds of things, from a change in lighting, to trees moving in the breeze and cobwebs on the camera lens.

It's estimated that operators review at least three alarms every minute, so if 95% of these are nuisance alarms, it makes it very difficult to concentrate on the real alarms that need to be dealt with. What if you could train a computer to filter out all those false alarms?

When it's been trained to look for human activity, AI is very powerful at reducing the burden of nuisance alarms on video monitoring staff. This is why machine learning is now so important to video analytics. Security companies are seeing the value that cutting-edge technology like AI can bring to their staff and their customers; increased speed and efficiency when dealing with real security threats.

Let's move on to how it's done. Training a computer to identify human activity in video footage is extremely difficult, which is why you need such advanced technology to make it work. Deep learning is capable of this task where other types of AI are not, because:

A. Neural networks (remember those?) are capable of processing complex information, and of learning how to improve those processes along the way

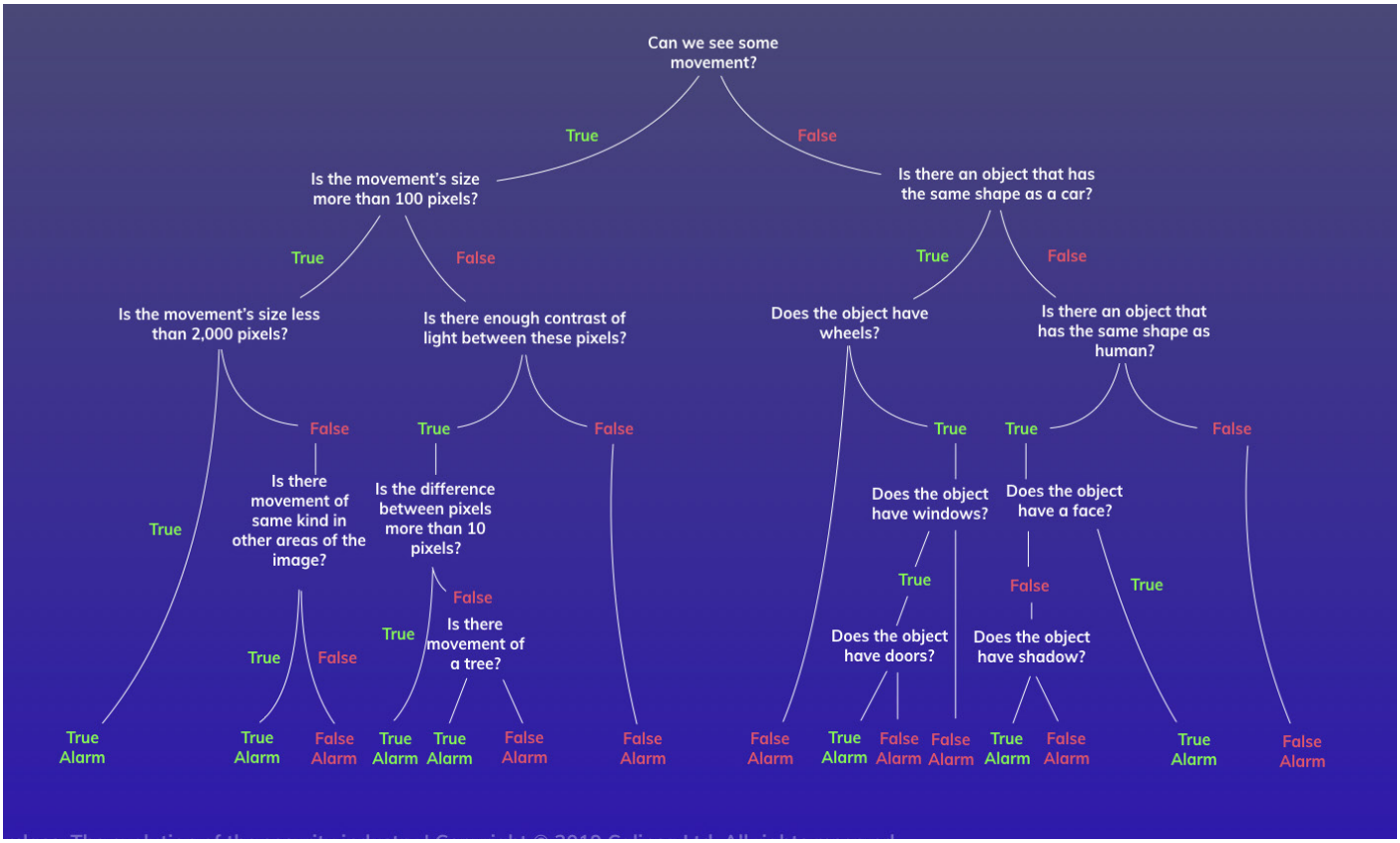
B. We're asking those neural networks to find patterns in images, which are a highly complex type of data

C. The technology now exists for neural networks to process such complex data at scale (i.e. reasonably quickly and affordably)

So deep learning needs three key elements to work: **algorithms, data and training on certain systems.** We'll look at how they come together to create intelligent video analytics.

ESSENTIAL 1: ALGORITHMS

The first essential of machine learning is **algorithms**. In its simplest form, algorithms are a set of rules to follow in order to solve a problem, such as a decision tree (see diagram).



For the past 10 years, a lot of video analytics software used **rule-based algorithms**: when confronted with an alarm, the software analysed it by following a pre-set decision tree of if-then rules set by an engineer. The problem is that humans aren't very good at designing these rules, especially for complex problems - there are simply too many possibilities in the data to consider.

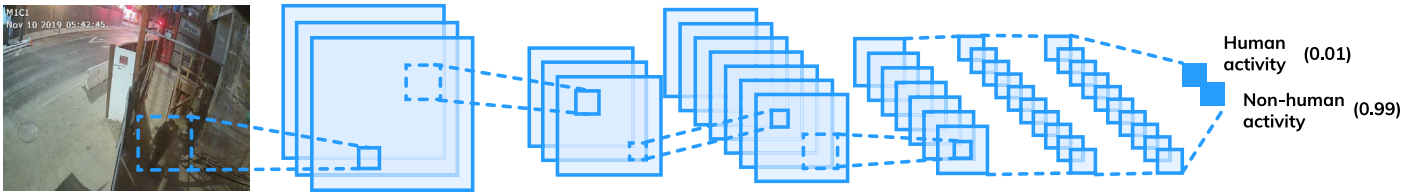
A better approach is to use machine learning so that these rules can be learned automatically from data. Although there are many machine learning algorithms available to use, the best ones for what we're trying to achieve in video analytics are based on **neural networks**, which are what we use at Calipsa.

The importance of neural networks

Back in the first section of this book, we looked briefly at neural networks and their importance in deep learning. Let's dive deeper into how they work and why they're so effective.

Unlike the decision tree method above, neural networks are not programmed with task-specific rules. Instead, multiple computational nodes are layered up and each of these nodes will analyse various elements in the image. Once the image has passed through the entire network (and been processed by all of the nodes), the final prediction is based on a much deeper analysis. Below, you can see how Calipsa's neural networks process an alarm.

In the first layers, the neural network recognises some basic features like edges and colours; as we move into the deeper layers, more advanced features are picked out. In the deepest layers, the machine's



representation of the image looks more and more abstract. The final output will predict whether the image contains human activity or not.

But doesn't the network still need to be told what to look for? Yes - but not in the same way as a rule-based algorithm. Neural networks use parameters, which are the guides that algorithms use to explain the data and predict outcomes in new data sets. Some parameters are set by humans (hyper-parameters, which are the overarching guides), and some are set by the machine itself as it learns.

Instead of changing each of the nodes' parameters manually, we start with a neural network in a random state and expose it to millions of images. As it processes the images, we allow the network to make a prediction and we update the parameters when it provides a wrong answer. Over time it learns little by little from its mistakes until it outperforms all the other methods, by learning how to get the optimal answer from the data it's been given. This is called supervised learning and the process of teaching the network is called training, which we will cover later.

Once a network has been trained, which for Calipsa is all about telling the difference between genuine and nuisance alarms, it can then be used to make predictions about new alarms.

This kind of learning is what makes neural networks so powerful: their ability to generalise what they have learned about alarms they've seen, and then apply those generalised concepts to alarms they have never seen before.

ESSENTIAL 2: DATA

Successful machine learning needs lots and lots of data. In most cases, that data needs to be labelled so the computer knows what to look for.

BBC / Public domain

To understand what labelled data is, let's go back to watching Planet Earth on Netflix. When you watch it, you're telling Netflix's algorithm that you would like it to recommend similar shows. But how does Netflix know what kind of a show Planet Earth is in the first place? Data labels. In their most basic form, they might look something like this:

Show title:	Planet Earth
Genres:	Documentary, Nature documentary, BBC documentary
Presenter:	David Attenborough
Released in:	2006

Based on this information, the algorithm can find you more nature documentaries, more shows by the BBC, more shows with David Attenborough, and so on. It will then use other viewers' choices to predict which of all these possible shows you are likely to watch.

Making images meaningful

When we're training a computer to analyse video footage, our data consists of images - these are two or three frames that contain a moving object. We have to label the images so the computer can learn which moving objects are human activity, and which are not. Over time, this helps it learn which alarms are true and which are false. So what does a label look like in an image?

As you can see, there is a green area around the person in the image. This green area is a data label, which tells the computer that everything inside the box is human activity: we draw green areas around people and cars when we're training a video analytics model.

Getting plenty of good labelled data is one of the biggest challenges in machine learning. At Calipsa, we regularly label a fraction of the alarm images we receive to continue training and improving our model; currently we label around 1000 alarms a day. This isn't done automatically - people have to manually label thousands of images so that we can be sure we're setting up the data correctly, ready for training.

Once we have used the labelled data to train our model, we use a second held-out set to test the model and see what it has learned. Running these tests regularly is how we can monitor the computer's progress and test its performance.

True or false? Testing machine intelligence

In video analytics, our main objective is to make the algorithm as accurate as possible; we want to identify genuine alarms, and to filter out nuisance alarms. When the computer makes a decision about an alarm, there are four possible outcomes, which enable us to measure performance:

True Positive **Image contains:** human activity **Machine identifies:** human activity **Outcome:** genuine alarm raised	**False Positive** **Image contains:** non-human activity **Machine identifies:** human activity **Outcome:** nuisance alarm raised
False Negative **Image contains:** human activity **Machine identifies:** non-human activity **Outcome:** genuine alarm ignored	**True Negative** **Image contains:** non-human activity **Machine identifies:** non-human activity **Outcome:** nuisance alarm ignored

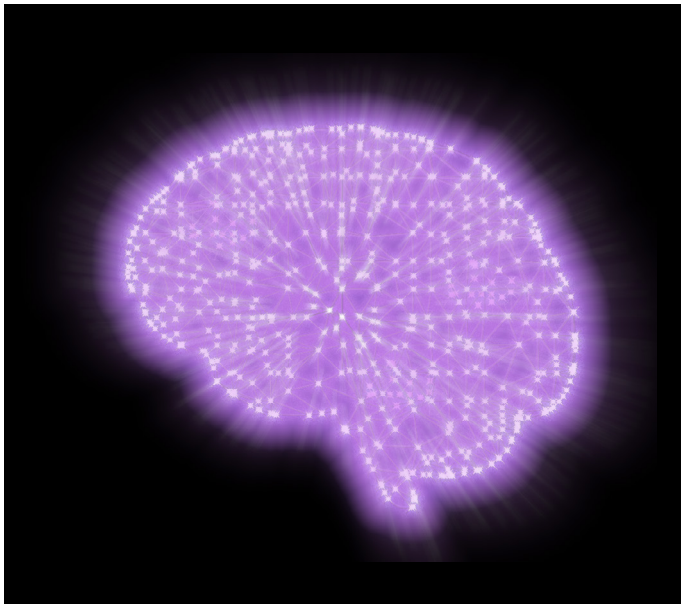
The two measures of success are known as **recall and reduction.** Recall is the percentage of human activity our model detects (true positives, or genuine alarms). Currently our recall rate is 99.95%. Reduction is the opposite: it's the percentage of non-human activity (true negatives, or nuisance alarms) our model detects. At the moment we have achieved a 92% reduction rate.

When we train our model, our goal is always to reduce our error rate and make our algorithms more accurate. This brings us back to the challenge of getting good labelled data. Your error rate is connected to the amount of data you have available: a good rule of thumb is that if you want to halve your error rate, you need to double the amount of data you put in.

While our technology is highly accurate, a big challenge is that we could always use more data. However, all of that data needs to be labelled before we can use it to teach the machine; that's thousands more images, all annotated by hand. So, even though machine learning can do incredible things, it still needs to be assisted by humans to point it in the right direction.

ESSENTIAL 3: TRAINING

Now that we've seen how important data is to machine learning, and why neural networks are so powerful, next let's look at the details of how to train it. The first thing to know is that training is time-consuming. Even with the most sophisticated technology in the world, teaching a machine to recognise objects in images is not easy.



Particularly challenging tasks, like image recognition, require networks that have a lot of parameters; for example, Calipsa's machine learning model has around 10 million parameters that help it understand what human activity looks like, all of which are set by the machine itself as it learns.

Faster training

The reason training is so time consuming is that there are a lot of parameters to estimate and lots of data is needed (the more, the better) so presenting each example to the network takes a lot of time. Fortunately, by using the right hardware, training can be reasonably quick.

One of the most common types of hardware used in machine learning is the **graphics processing unit (GPU).** GPUs are specifically designed to perform thousands of simple mathematical operations in parallel. Advances in GPUs in the early 2010s meant that rapid, complex processing of images was finally possible at scale.

It currently takes approximately a day to train our model at Calipsa. Given the amount of data we expose the network to, this is a reasonable amount of time for training to take. If we wanted to reduce the time it takes, then we would need to use more hardware, or even higher spec hardware.

This is a challenge many machine learning teams face: finding the balance between efficiency and cost. The more hardware you use, the better your training model. However, the more hardware you use, the more expensive it is to run!

The challenges

One of the reasons we need to use so much data in machine learning is to avoid a problem called **overfitting**. Overfitting is where a machine learning model memorises the training data, so it can only make predictions based on a specific set of situations; for example, it can only spot human activity if it matches existing examples.

Instead, we want the network to learn a general representation of human activity, so that it can make sensible decisions in new and unforeseen situations. We always use one set of data for training, and a second, unseen set for testing afterwards. Ultimately, avoiding overfitting is why we need to use so much data, because we have so many parameters in our model.

As well as using multiple data sets, we also put constraints on the parameters to avoid them becoming too specific. This process is known as **regularisation.** In machine learning, and particularly in video analytics, our goal is to teach the machine to recognise generalised concepts in very rich, complex data.

To be able to pick out a person in an otherwise very rich scene full of plants, buildings and animals, the neural network needs to strike a balance between learning what a human looks like, but not going on to learn every single feature of the image. Just like decision trees, overfitted parameters can become inflexible (and consequently inaccurate) when they don't have enough data to work with, or if the quantity and relevance of those parameters goes unchecked.

Having too many parameters can raise other issues besides overfitting. Neural networks are complex and layered, and while this is one of their benefits, it can also become a drawback. One of these challenges is the **vanishing gradient problem**.

When we send information through a neural network, it passes through multiple layers of parameters. The more complex the network becomes, the more layers there are to pass through. As it gets sent back through all the preceding layers, the signal can get lost and the network stops learning. This is another reason why we might want to "prune" our parameters and make sure they don't over-complicate our model.

However, as with data and hardware, there is a balance to be struck to achieve the best result. We need lots of parameters to make sure the algorithm is sophisticated enough to understand the data - but not so many that it hinders the machine's learning process.

THE FUTURE?

Machine learning in video analytics can achieve different goals, which loosely fall into three categories:

- **Classification**: identifies objects contained in an image
- **Detection**: specifies the location(s) of multiple objects in an image
- **Segmentation**: identifies different groups of pixels in an image as objects of various shapes and sizes

At Calipsa, our False Alarm Filtering Platform does a combination of these things to identify genuine alarms: our model can classify what a human or car looks like, pick out its location in an image, and recognise the general shape and size of these moving objects across two or more images.

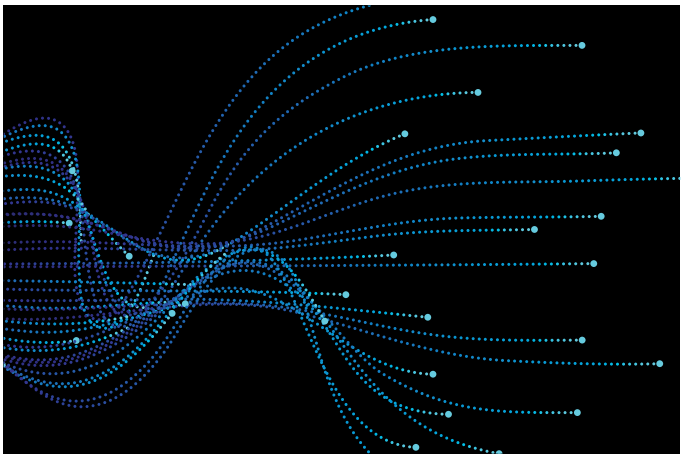
Now, our model can very accurately identify people and cars. In time, as we continue to train our software and as we see new developments in computer vision, it's likely that video analytics will become even more sophisticated.

The next big step in video analytics will be identifying behaviours to predict criminal intent. Now that the model can identify human activity, crime prevention could be pushed even further by learning what movement and behaviours are most likely to lead to criminal activity.

To do this, the machine's understanding of all the elements in an image will have to become even more sophisticated: as well as knowing what a person looks like, it will need to identify all the things that a person interacts with in their environment.

These developments raise even more challenges and questions, both technical and ethical: how do we get a machine to recognise human behaviour, and do the benefits outweigh the ethical implications of this level of surveillance?

CONCLUSION



When machine learning has the right data, training and algorithms to support it, it can do its job - which is to look for patterns. Until the early 2010s, pattern recognition didn't work so well with images. In ten years, technology has come a long way and it's now much easier for us to train computers to spot patterns in images and video.

However, as we hope this book has explained, machine learning still requires a lot of input from people to make it truly effective, and there are many challenges that engineers encounter along the way.

Now that we have this technology at our fingertips, it has helped us to make huge leaps forward in crime prevention: our False Alarm Filtering Platform makes highly accurate decisions that increase monitoring operators' speed and efficiency when dealing with genuine alarms. What's more, the platform continuously learns and improves over time, so it's always getting better at spotting human activity.

Machine learning enables us to do things better and faster than we could have done before, and over time it's only going to improve. In our opinion, it's going to be a big part of our lives for the foreseeable future.

GLOSSARY

Algorithms - A set of rules a computer will follow in order to solve a problem.

Artificial intelligence (AI) - Any technique that enables a computer to mimic human intelligence.

Classification - When a neural network identifies objects contained in an image.

Data labels - Labels manually added to data in order to help a neural network identify patterns.

Deep learning - A subset of machine learning where the algorithm trains itself to perform a task by exposing deep neural networks to vast amounts of data.

Detection - When a neural network specifies the location(s) of multiple objects in an image.

Genuine alarms - Security camera footage that contains human activity in a scene where people aren't supposed to be. Also known as a True positive (see page 7).

Graphics processing unit (GPU) - One of the most common types of hardware used in machine learning, GPUs are specifically designed to perform thousands of simple mathematical operations in parallel.

If-then rules - Conditional rules set by an engineer for a computer to follow, often used in rule-based algorithms.

Machine learning - A subset of artificial intelligence, machine learning algorithms use statistics to find patterns in vast amounts of data, applying those patterns to make predictions.

Neural networks - A computing system made up of many layers of computational nodes, giving it an enhanced ability to spot patterns and process much larger amounts of complex information.

Nuisance alarms - Security camera footage that contains non-human activity, but is still sent to operators as a security alert. Also known as a False positive (see page 7).

Overfitting - When a machine learning model memorises the training data, so it can only make predictions based on a specific set of situations.

Parameters - The guides that algorithms use to explain the data and predict outcomes in new data sets.

Recall - The percentage of true positives a machine learning model detects.

Reduction - The percentage of true negatives a machine learning model detects.

Regularisation - Putting constraints on parameters to avoid them becoming too specific, with an aim to help the computer learn more generalised concepts.

Rule-based algorithms - Algorithms that find patterns in data, by following rules that have been pre-set by a human.

Segmentation - When a neural network identifies different groups of pixels in an image as objects of various shapes and sizes.

Supervised learning - Training a neural network by exposing it to data, and correcting its mistakes. Over time it learns from these mistakes until it learns how to get the optimal answer from the data it's been given.

Training - The process of teaching a neural network how to solve a problem and/or learn a generalised concept.

Vanishing gradient problem - When information has to pass through an increasingly complex network, the signal can get lost, causing the network to stop learning.

ABOUT AUTHOR

David Hall

Head of Machine Learning - Calipsa

David joined Calipsa in January to lead our Machine Learning team, where he creates, develops and implements the AI systems that enable Calipsa to detect and prevent crime. Prior to joining Calipsa, he was a freelance consultant providing custom AI solutions for clients with a focus on vision applications. He brings 10 years' experience in machine learning and computer vision, including a PhD in Computational Vision from Caltech, and numerous published articles on different areas of visual recognition.