WHITEPAPER

KI in der Videoanalyse

Überlegungen zur Analytik auf der Grundlage von maschinellem Lernen und Deep Learning

März 2021

AXIS COMMUNICATIONS

Inhalt

1 Zusammenfassung

KI-basierte Videoanalyse gehört zu den meistdiskutierten Themen in der Videoüberwachungsindustrie. Manche der Anwendungen können die Datenanalyse deutlich beschleunigen und sich wiederholende Aufgaben automatisieren. Aber die heutigen KI-Lösungen können die Erfahrungen und Fähigkeiten zur Entscheidungsfindung des menschlichen Bedieners nicht ersetzen. Ihre Stärke liegt vielmehr in der Kombination von KI-Lösungen mit einem menschlichen Bediener, um dessen Effizienz zu steigern.

Das AI-Konzept umfasst Algorithmen für maschinelles Lernen und für Deep Learning. Beide Typen erstellen automatisch aus großen Mengen von Probedaten *(Trainingsdaten)* ein mathematisches Modell, um Ergebnisse ohne spezielle Programmierung berechnen zu können. Ein KI-Algorithmus wird durch einen iterativen Prozess entwickelt, bei dem ein Zyklus aus dem Sammeln von Trainingsdaten, dem Beschriften von Trainingsdaten, dem Verwenden der beschrifteten Daten zum Trainieren des Algorithmus und dem Testen des trainierten Algorithmus so lange wiederholt wird, bis die gewünschte Qualitätsstufe erreicht ist. Danach ist der Algorithmus bereit für den Einsatz in einer kommerziellen Analyseanwendung, die an einem Überwachungsstandort installiert werden kann. Damit ist das Training abgeschlossen und die Anwendung kann nichts Neues mehr lernen.

Eine typische Aufgabe für eine KI-basierte Videoanalyse ist die visuelle Erkennung und Unterscheidung von Menschen und Fahrzeugen in einem Videostream. Ein Algorithmus für *maschinelles Lernen* hat die Kombination visueller Merkmale gelernt, die diese Objekte definieren. Ein Algorithmus für *Deep Learning* ist viel ausgefeilter und kann – sofern er entsprechend trainiert ist – viel komplexere Objekte erkennen. Dies erfordert jedoch einen deutlich höheren Entwicklungs- und Trainingsaufwand und viel mehr Rechenressourcen bei der Ausführung der fertigen Anwendung. Für genau spezifizierte Überwachungsaufgaben muss daher überlegt werden, ob eine dedizierte, optimierte Anwendung mit maschinellem Lernen nicht vielleicht ausreichen würde.

Dank der Entwicklung von Algorithmen und der immer größeren Rechenleistung der Kameras ist jetzt eine erweiterte KI-basierte Videoanalyse direkt in der Kamera (Peripherie) möglich. Das ermöglicht bessere Echtzeit-Funktionen, weil die Anwendungen direkt auf das unkomprimierte Videomaterial zugreifen können. Mit einer dedizierten Hardware-Beschleunigung wie MLPU (Machine Learning Processing Unit) und DLPU (Deep Learning Processing Unit) in den Kameras kann Edge-basierte Analyse stromsparender als bei einer CPU oder einem GPU (Grafikprozessor) umgesetzt werden.

Vor der Installation einer KI-basierten Videoanalyseanwendung müssen die Empfehlungen des Herstellers auf der Grundlage bekannter Vorbedingungen und Einschränkungen genau studiert und befolgt werden. Jede Überwachungsanwendung ist einzigartig, und die Leistung der Anwendung muss direkt am jeweiligen Standort untersucht werden. Bleibt die Qualität hinter den Erwartungen zurück, muss die gesamte Anlage untersucht werden, nicht nur die Analyseanwendung. Die Möglichkeiten der Videoanalyse sind von vielen Faktoren abhängig: der Kamera-Hardware, Kamerakonfiguration, Videoqualität, Szenendynamik und Beleuchtung. Oft lässt sich die Qualität der Videoanalytik in der Anlage verbessern, indem man sich überlegt, wie sich diese Faktoren auswirken, und sie entsprechend optimiert.

Die KI wird immer häufiger in der Überwachung eingesetzt und bietet unbestrittene Vorteile für die betriebliche Effizienz und neue Einsatzmöglichkeiten. Doch muss immer sorgfältig abgewogen werden, wann und wo sich ihr Einsatz wirklich lohnt.

2 Einführung

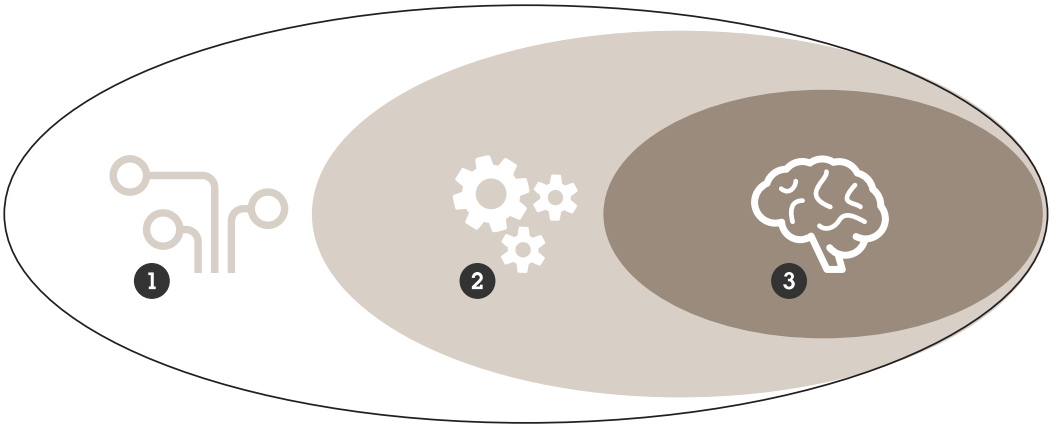
KI, künstliche Intelligenz, wird seit der Erfindung der ersten Computer entwickelt und diskutiert. Die revolutionärsten Verkörperungen sind zwar noch nicht Realität geworden, aber KI-basierte Technologien werden heute häufig für präzise Aufgaben in Anwendungen wie Spracherkennung, Suchmaschinen oder virtuellen Assistenten eingesetzt. KI kommt außerdem immer öfter im Gesundheitswesen zum Einsatz, wo sie wertvolle Ressourcen beispielsweise bei der Röntgendiagnose oder Analyse von Netzhautscans liefert.

KI-basierte Videoanalyse gehört zu den meistdiskutierten Themen in der Videoüberwachungsindustrie, die sich viel von ihr verspricht. Es gibt Anwendungen auf dem Markt, die KI-Algorithmen nutzen, um Datenanalysen erfolgreich zu beschleunigen und sich wiederholende Aufgaben zu automatisieren. Aber im allgemeineren Überwachungskontext sollte die KI heute und in naher Zukunft nur als eines von mehreren Elementen zum Aufbau zuverlässiger Lösungen gesehen werden.

Dieses White Paper liefert den technologischen Hintergrund zu den Algorithmen für maschinelles Lernen und Deep Learning und wie sie für die Videoanalyse entwickelt und angewendet werden. Dabei gehen wir kurz auf die KI-Beschleunigungshardware sowie die Vor- und Nachteile einer KI-basierten Analytik in der Peripherie im Vergleich zu einem Server ein. Außerdem untersuchen wir unter Berücksichtigung verschiedener Faktoren, wie sich die Voraussetzungen für KI-basierte Videoanalyse optimieren lassen.

3 KI, maschinelles Lernen und Deep Learning

Künstliche Intelligenz (KI) ist ein breit gefasster Begriff, der Maschinen bezeichnet, die komplexe Aufgaben ausführen können und Anzeichen von Intelligenz aufzuweisen scheinen. Deep Learning und maschinelles Lernen sind Teilbereiche der KI.



1 Künstliche Intelligenz
2 Maschinelles Lernen
3 Deep Learning

3.1 Maschinelles Lernen

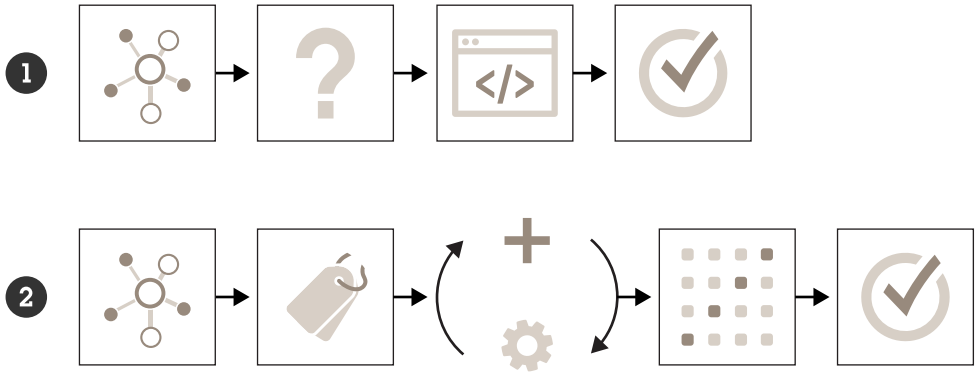
Maschinelles Lernen ist ein Teilbereich der KI, bei dem mithilfe von statistischen Lernalgorithmen Systeme aufgebaut werden, die während einer Trainingsphase ohne gezielte Programmierung automatisch lernen und sich verbessern.

In diesem Abschnitt unterscheiden wir zwischen herkömmlicher Programmierung und maschinellem Lernen im Kontext der *Computer Vision* – einer Wissenschaft, mit der Computer durch Analyse von Bildern oder Videos verstehen lernen sollen, was in einer Szene geschieht.

Herkömmlich programmierte Computer Vision basiert auf Verfahren zur Berechnung der *Merkmale* eines Bildes. So suchen die Computer zum Beispiel nach ausgeprägten Kanten und Eckpunkten. Diese Funktionen müssen manuell von einem Algorithmus-Entwickler programmiert werden, der weiß, auf welche Teile der Bilddaten es ankommt. Daraufhin kombiniert er diese Funktionen so, dass der Algorithmus auf den Inhalt der Szene schließen kann.

Die Algorithmen für maschinelles Lernen bauen automatisch aus großen Mengen von Probedaten *(Trainingsdaten)* ein mathematisches Modell auf, damit der Computer lernt, Entscheidungen durch Berechnung der Ergebnisse zu fällen, ohne speziell dafür programmiert zu sein. Die Merkmale werden weiterhin von einem Menschen programmiert, aber der Algorithmus lernt anhand großer Mengen von gelabelten oder *annotierten* Trainingsdaten, wie er diese Merkmale kombinieren soll. In diesem White Paper bezeichnen wir diese Technik der Verwendung handgefertigter Merkmale in gelernten Kombinationen als *klassisches maschinelles Lernen.*

Anders gesagt: Bei maschinellem Lernen muss der Computer geschult werden, damit man das gewünschte Programm erhält. Die Daten werden von Menschen gesammelt und mit Anmerkungen versehen, manchmal unterstützt durch von Servercomputern vergebene Annotationen. Das Ergebnis wird so lange in das System eingespeist, bis der Computer genug gelernt hat, um das gewünschte Ziel zu erkennen, wie zum Beispiel eine bestimmte Art von Fahrzeug. Das geschulte Modell wird zum Programm. Sobald das Programm fertiggestellt ist, lernt das System nichts mehr dazu.



1 *Traditionelle Programmierung:*
Daten werden gesammelt. Programmkriterien werden definiert. Das Programm wird (von einem Menschen) geschrieben. Fertig.
2 *Maschinelles Lernen:*
Daten werden gesammelt. Die Daten werden gelabelt. Das Modell wird einem iterativen Training unterzogen. Das fertig geschulte Modell wird zum Programm. Fertig.

Der Vorteil der KI gegenüber der traditionellen Programmierung beim Aufbau eines Computer-Vision-Programms ist ihre Fähigkeit, große Datenmengen zu verarbeiten. Ein Computer kann Tausende von Bildern durchsehen, ohne dass seine Konzentration nachlässt, während ein menschlicher Programmierer nach einer gewissen Zeit ermüdet und unkonzentriert wird. In dieser Hinsicht kann eine KI-Anwendung sehr viel genauer sein. Je komplizierter aber die Anwendung, desto schwerer wird es für die Maschine, das gewünschte Ergebnis zu liefern.

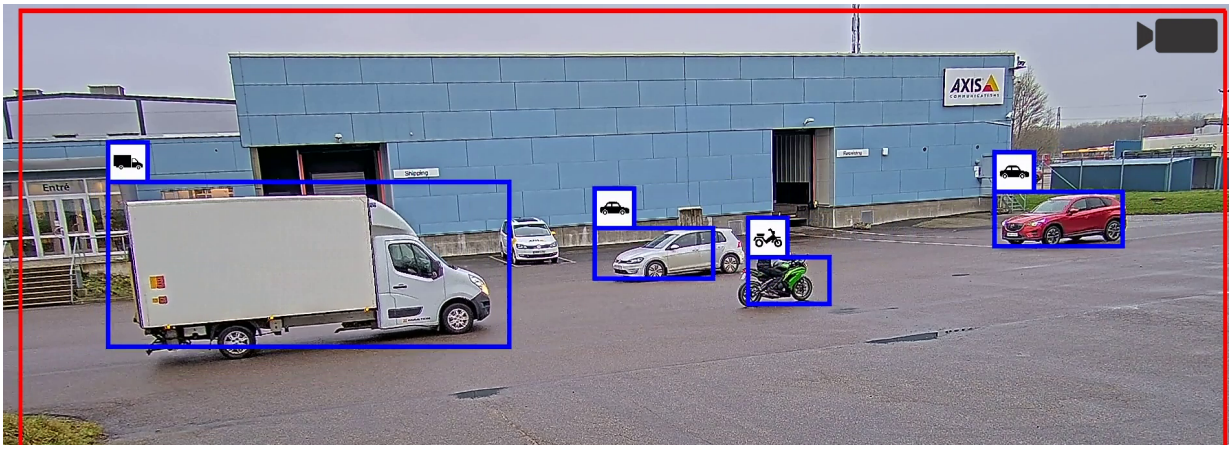
3.2 Deep Learning

Deep Learning ist eine verfeinerte Version von maschinellem Lernen, wobei sowohl die Extraktion als auch die Kombination der Merkmale in umfangreichen Regelstrukturen datengesteuert erlernt werden. Der Algorithmus kann automatisch festlegen, nach welchen Merkmalen er in den Trainingsdaten suchen soll. Außerdem kann er sehr tiefe Strukturen verketteter Kombinationen von Merkmalen lernen.

Die bei Deep Learning angewendeten Algorithmen sind an die Funktionsweise von Neuronen angelehnt. Wie das Gehirn kombinieren sie die Neuronenausgaben in einer mehrstufigen Hierarchie, einem *Netzwerk* miteinander verketteter Regeln, zu höherem Wissen. Im menschlichen Gehirn bestehen sogar die Kombinationen selbst aus Neuronen, so dass es keine Unterscheidung zwischen der Extraktion und der Kombination von Merkmalen gibt. Beide sind in gewissem Sinne identisch. Wissenschaftler haben diese Strukturen in so genannten *künstlichen neuronalen Netzen* simuliert, dem am häufigsten verwendeten Typ von Algorithmus bei Deep Learning. Eine kurze Übersicht über neuronale Netze finden Sie im Anhang dieses Dokuments.

Die Algorithmen von Deep Learning ermöglichen den Aufbau komplexer visueller Detektoren, die auf die automatische Erkennung extrem komplexer Objekte trainiert werden können, unabhängig von Maßstab, Drehung und anderen Abweichungen.

Diese Flexibilität ist möglich, weil Deep Learning-Systeme aus einer viel größeren Datenmenge und sehr viel unterschiedlicheren Daten lernen können als es beim klassischen maschinellen Lernen üblich ist. In den meisten Fällen sind sie deutlich leistungsfähiger als vom Menschen programmierte Computer-Vision-Algorithmen. Deshalb eignet sich Deep Learning besonders gut für komplexe Probleme, bei denen die Merkmale nur schwer von menschlichen Experten kombiniert werden können, wie Bildklassifizierung, Sprachverarbeitung und Objekterkennung.



Objekterkennung auf der Grundlage von Deep Learning kann komplexe Objekte klassifizieren. In diesem Beispiel kann die Analytik nicht nur Fahrzeuge erkennen, sondern diese auch klassifizieren.

3.3 Klassisches maschinelles Lernen vs. Deep Learning

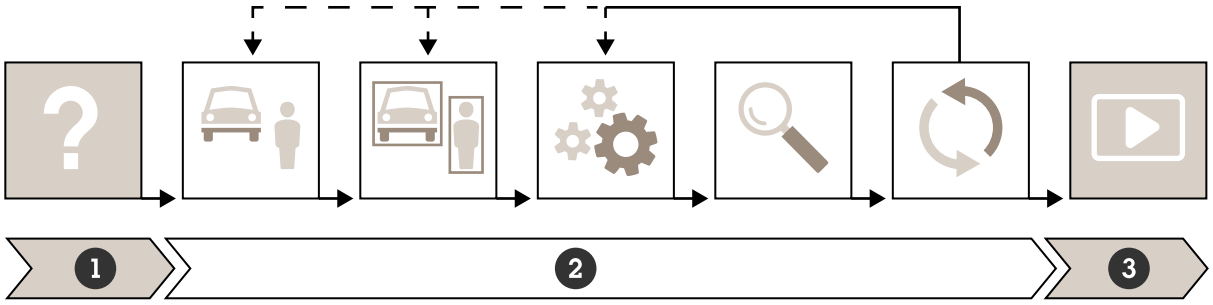
Beide arbeiten zwar mit ähnlichen Algorithmen, aber Deep Learning erfordert in der Regel einen viel größeren Satz erlernter Merkmalskombinationen als klassisches maschinelles Lernen. Auf Deep Learning basierende Analytik kann also flexibler sein und – sofern entsprechend geschult – viel komplexere Aufgaben ausführen.

Für die spezielle Analytik in der Überwachung könnte unter Umständen auch ein spezieller, optimierter klassischer Algorithmus für maschinelles Lernen durchaus ausreichen. In klar umrissenen

Aufgabenbereichen kann dieser ähnliche Ergebnisse liefern wie ein Deep Learning-Algorithmus, erfordert aber weniger Rechenoperationen und kann deshalb kostengünstiger und stromsparender sein. Außerdem erfordert er viel weniger Trainingsdaten, was den Entwicklungsaufwand drastisch reduziert.

4 Die Stadien des maschinellen Lernens

Die Entwicklung eines Algorithmus für maschinelles Lernen besteht aus einer Reihe von Schritten und Iterationen (unten eine grobe Veranschaulichung), bis die fertige Analyseanwendung eingerichtet werden kann. Den Kern der Analyseanwendung bilden ein oder mehrere Algorithmen, zum Beispiel zur Objekterkennung. Bei Algorithmen, die auf Deep Learning beruhen, ist der Kern des Algorithmus das Deep-Learning-Modell.



1 *Vorbereitung: Festlegung des Anwendungszwecks.*
2 *Training: Sammeln von Trainingsdaten. Versehen der Daten mit Anmerkungen (Annotieren). Trainieren des Modells. Testen des Modells. Ist die Qualität nicht wie erwartet, werden die vorangegangenen iterativen Verbesserungsschritte wiederholt.*
3 *Deployment: Installation und Verwendung der fertigen Anwendung.*

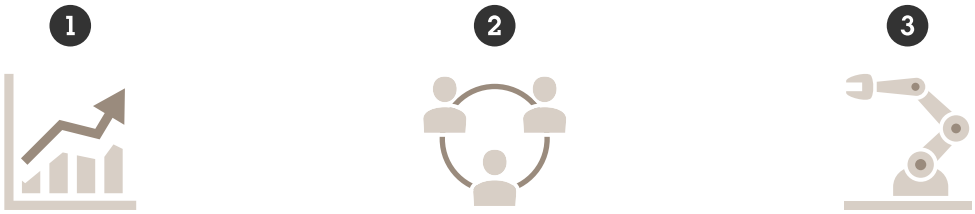
4.1 Datenerfassung und Annotation

Die Entwicklung einer KI-basierten Analyseanwendung erfordert riesige Datenmengen. In der Videoüberwachung sind dies in der Regel Bilder und Videoclips von Personen und Fahrzeugen oder anderer Zielobjekte. Um die Daten für eine Maschine oder einen Computer erkennbar zu machen, müssen die Daten mit Annotationen versehen werden, in denen die relevanten Objekte kategorisiert und gelabelt (mit einem Etikett versehen) werden. Die Datenannotation ist ein überwiegend manueller, arbeitsaufwändiger Prozess. Die vorbereiteten Daten müssen eine ausreichend große Auswahl von für den Kontext relevanten Beispielen abdecken, in dem die Analyseanwendung eingesetzt werden soll.

4.2 Schulungen

Bei der Schulung (dem Anlernen) werden dem Modell annotierte Daten vorgelegt. In einem iterativen Trainings-Framework wird das Modell so lange immer weiter angepasst und verbessert, bis die gewünschte

Qualität erreicht ist. Das Modell wird also für die Lösung einer vorgegebenen Aufgabe optimiert. Für das Training werden hauptsächlich drei Methoden eingesetzt.



1 *Überwachtes Lernen: Das Modell lernt, richtige Vorhersagen zu treffen*
2 *Selbstständiges Lernen: Das Modell lernt, Cluster zu erkennen*
3 *Bestärkendes Lernen: Das Modell lernt aus Fehlern*

4.2.1 Überwachtes Lernen

Überwachtes Lernen ist die heute gebräuchlichste Methode für maschinelles Lernen. Sie kann als Lernen anhand von Beispielen beschrieben werden. Die Trainingsdaten werden mit klaren Annotationen versehen. Die Eingabedaten sind also bereits mit dem gewünschten Ausgabeergebnis verknüpft.

Überwachtes Lernen erfordert prinzipiell eine große Menge annotierter Daten, und die Leistung des trainierten Algorithmus ist direkt von der Qualität der Trainingsdaten abhängig. Der wichtigste Qualitätsaspekt ist, dass die Trainingsdatensätze alle möglichen Eingabedaten aus einer echten Einsatzsituation berücksichtigen müssen. Bei der Objekterkennung muss der Entwickler den Algorithmus mit vielen unterschiedlichen Bildern, verschiedenen Objektinstanzen, Ausrichtungen, Maßstäben, Beleuchtungssituationen, Hintergründen und Ablenkungen trainieren. Nur mit repräsentativen Trainingsdaten für den geplanten Einsatzzweck kann die spätere Analyseanwendung auch bei neuen Daten, die sie während der Trainingsphase nicht gesehen hat, zuverlässige Voraussagen machen.

4.2.2 Selbstständiges Lernen

Beim selbstständigen Lernen werden anhand von Algorithmen neue Datensätze analysiert und in Gruppen eingeteilt. Dies ist keine übliche Trainingsmethode in der Überwachungsbranche, weil das Modell umfangreiche Kalibrierungen und Tests erfordert und die Qualität trotzdem nicht vorhersehbar sein kann.

Die Datensätze müssen für die Analyseanwendung relevant sein, aber sie brauchen nicht deutlich gelabelt oder markiert zu werden. Zwar fallen die manuellen Annotationen weg, doch dafür muss die erforderliche Menge an Bildern oder Videos für die Schulung um ein Vielfaches erhöht werden. Während der Trainingsphase identifiziert das trainierte Modell – unterstützt durch das Trainings-Framework – gemeinsame Merkmale in den Datensätzen. So kann es während der Deployment-Phase Daten nach Mustern zu Gruppen zusammenstellen, aber auch Anomalien erkennen, die in keine der gelernten Gruppen passen.

4.2.3 Bestärkendes Lernen

Bestärkendes Lernen wird beispielsweise in der Robotik, Industrieautomatisierung und Unternehmensstrategieplanung eingesetzt, aber weil die Methode umfangreiches Feedback benötigt, ist sie bis heute nur eingeschränkt im Überwachungsbereich einsetzbar. Beim bestärkenden Lernen geht es darum, angemessene Maßnahmen zu ergreifen, um den möglichen *Gewinn* in einer spezifischen Situation zu maximieren. Der Gewinn wird größer, wenn das Modell die richtigen Entscheidungen trifft. Dieser Algorithmus nutzt keine Daten-Label-Paare für das Training, sondern wird stattdessen durch das Testen der

Entscheidungen in Interaktion mit der Umgebung bei gleichzeitigem Messen des Gewinns optimiert. Dieser Algorithmus zielt darauf ab, eine Strategie für Aktionen zu erlernen, die den Gewinn maximiert.

4.3 Test

Wenn das Modell trainiert ist, muss es umfassend getestet werden. Dieser Schritt besteht in der Regel aus einem automatischen Teil, ergänzt durch umfangreiche Tests in realen Einsatzsituationen.

Im automatisierten Teil erhält die Anwendung Bezugspunkte in Form neuer Datensätze, die das Modell während des Trainings noch nicht gesehen hat. Sind diese Bezugspunkte nicht zufriedenstellend, beginnt der Ablauf von vorn: Neue Trainingsdaten werden gesammelt, Anmerkungen hinzugefügt oder angepasst und das Modell neu trainiert.

Nachdem die gewünschte Qualität erreicht ist, beginnt ein Feldtest. Darin wird die Anwendung mit echten Szenarien konfrontiert. Deren Menge und Schwankungsbreite ist vom Umfang des späteren Einsatzzwecks abhängig. Je geringer der Umfang, desto weniger Variationen müssen getestet werden. Je größer der Umfang, desto mehr Tests sind erforderlich.

Die Ergebnisse werden wiederum verglichen und ausgewertet. Auch nach diesem Schritt kann der ganze Ablauf noch einmal von neuem beginnen. Ein weiteres mögliches Ergebnis könnte die Festlegung von Vorbedingungen sein, eine Beschreibung eines bekannten Szenarios, für das die Anwendung nicht oder nur eingeschränkt empfohlen wird.

4.4 Deployment

Die Deployment-Phase wird auch als Schlussfolgerungs- oder Vorhersagephase bezeichnet. Die Anwendung eines trainierten maschinellen Lernmodells liefert *Schlussfolgerungen* oder *Vorhersagen.* Der Algorithmus nutzt das in der Trainingsphase Gelernte, um das gewünschte Ergebnis zu erzeugen. Im Kontext der Analyse für Überwachungszwecke ist die Schlussfolgerungsphase die Anwendung, die in einem Überwachungssystem ausgeführt wird, um reale Szenen zu überwachen.

Um mit einem auf maschinellem Lernen basierenden Algorithmus mit Audio- oder Video-Eingabedaten in Echtzeit hohe Leistung zu erzielen, ist normalerweise eine Hardwarebeschleunigung notwendig.

5 Edge-basierte Analyse

Leistungsfähige Videoanalytik war bisher serverbasiert, weil sie mehr Strom und Kühlung erforderte als eine Kamera zur Verfügung stellen konnte. Weiterentwicklungen bei den Algorithmen und die immer höhere Rechenleistung von Edge-Geräten in den letzten Jahren ermöglichen heute eine leistungsfähige KI-Videoanalyse in der Peripherie.

Edge-basierte Analyseanwendungen bieten klare Vorteile: Sie haben mit sehr geringer Latenz Zugriff auf das unkomprimierte Videomaterial, so dass Echtzeitanwendungen ohne die zusätzlichen Kosten und Komplexität der Verschiebung der Daten in die Cloud für die Verarbeitung möglich werden. Edge-basierte Analytik ist außerdem mit weniger Hardware- und Deployment-Kosten verbunden, da das Überwachungssystem weniger Serverressourcen erfordert.

Manche Anwendungen könnten von einer Kombination aus Edge-basierter und serverbasierter Verarbeitung profitieren, mit Vorverarbeitung in der Kamera und Weiterverarbeitung im Server. Ein solches Hybridsystem kann eine kosteneffiziente Skalierung von Analyseanwendungen durch die Arbeit in mehreren Kamerastreams ermöglichen.

6 Hardwarebeschleunigung

Viele Analyseanwendungen können auf unterschiedlichen Plattformen ausgeführt werden, aber bei begrenzter Stromversorgung erzielt man mit einer gezielten Hardwarebeschleunigung ein viel besseres Ergebnis. Hardwarebeschleunigung ermöglicht eine stromsparende Implementierung von Analyseanwendungen. Sie kann gegebenenfalls durch Server- und Cloud-Rechenressourcen ergänzt werden.

- **GPU (Grafikprozessor):** GPUs wurden vorrangig für Grafikverarbeitungsanwendungen entwickelt, aber sie werden auch zur KI-Beschleunigung in Server- und Cloud-Plattformen eingesetzt. Obwohl sie manchmal auch in eingebetteten Systemen (Edge) eingesetzt werden, sind GPUs im Hinblick auf ihren Strombedarf nicht ideal, um Schlussfolgerungen bei maschinellem Lernen zu ziehen.
- **MLPU (Machine Learning Processing Unit):** MLPUs können spezifische klassische Algorithmen für maschinelles Lernen beschleunigen, um Aufgaben der Computer Vision mit hoher Stromeffizienz auszuführen. Sie wurden für die Objekterkennung in Echtzeit bei nur wenigen gleichzeitigen Objekttypen entwickelt, wie beispielsweise Personen und Fahrzeuge.
- **DLPU (Deep Learning Processing Unit):** Kameras mit eingebautem DLPU können allgemeine Algorithmen für Deep Learning sehr stromsparend beschleunigen, was gezieltere Objektklassifizierungen erlaubt.

7 Die KI befindet sich noch in der frühen Entwicklungsphase

Es ist verlockend, das Potential einer KI-Lösung mit den Möglichkeiten eines Menschen zu vergleichen. Ein Mensch hat bei der Videoüberwachung nur eine kurze Aufmerksamkeitsspanne, während ein Computer schnell große Datenmengen verarbeiten kann, ohne jemals zu ermüden. Es wäre jedoch ein grundlegendes Missverständnis anzunehmen, dass KI-Lösungen den menschlichen Bediener vollständig ersetzen könnten. Die besten Ergebnisse lassen sich vielmehr durch eine realistische Kombination erzielen, bei der KI-Lösungen den menschlichen Bediener unterstützen und seine Effizienz steigern.

Oft wird behauptet, maschinelles Lernen oder Deep-Learning-Lösungen könnten automatisch oder durch Erfahrung lernen. Heute verfügbare KI-Systeme können nach der Installation *nicht* automatisch neue Fertigkeiten erlernen und sich *nicht* an Geschehnisse erinnern. Um die Leistung des Systems zu verbessern, muss es während der überwachten Lernsitzungen mit besseren und genaueren Daten neu trainiert werden. Unüberwachtes Lernen erfordert in der Regel große Datenmengen zur Erzeugung von Clustern, weshalb es in Videoüberwachungsanwendungen nicht eingesetzt wird. Stattdessen wird es vorrangig zur Analyse großer Datensätze eingesetzt, um Anomalien, zum Beispiel bei Finanztransaktionen, zu finden. Die meisten der als „selbstlernend" beschriebenen Strategien in der Videoüberwachung basieren auf statistischen Datenanalysen und nicht auf dem tatsächlichen Einsatz von Deep-Learning-Modellen.

Menschliche Erfahrung ist vielen KI-basierten Analyseanwendungen für Überwachungszwecke nach wie vor überlegen. Das gilt insbesondere bei der Ausführung sehr allgemeiner Aufgaben, bei denen das Kontextverständnis entscheidend ist. Eine auf maschinellem Lernen basierende Anwendung könnte eine „rennende Person" erfolgreich erkennen, wenn sie speziell dafür geschult ist, aber im Gegensatz zu einem Menschen, der die Daten im Kontext sieht, erkennt die Anwendung nicht, warum die Person rennt – um den Bus zu erwischen oder um dem Polizisten in der Nähe zu entkommen? Trotz aller Versprechungen von Unternehmen, die KI in ihren Analyseanwendungen für die Überwachung nutzen, kann die Anwendung noch lange nicht mit dem gleichen Verständnis wie ein Mensch begreifen, was sie im Video sieht.

Aus diesem Grund können KI-basierte Analyseanwendungen auch falsche oder keine Alarme auslösen. Dies passiert am wahrscheinlichsten in komplexen Umgebungen mit viel Bewegung. Es könnte aber auch

eine Person sein, die einen großen Gegenstand trägt. Dieser macht die Merkmale des Menschen für die Anwendung unkenntlich und reduziert die Wahrscheinlichkeit für eine richtige Klassifizierung.

KI-basierte Analytik sollte heute als Unterstützung eingesetzt werden, um zum Beispiel die Relevanz eines Vorfalls grob einzuschätzen, bevor eine menschliche Wachperson alarmiert wird, die über das weitere Vorgehen entscheidet. Auf diese Weise sorgt die KI für Skalierbarkeit, und der menschliche Bediener beurteilt die Situation.

8 Überlegungen für eine optimale Analyse-Performance

Zur Steuerung der Qualitätserwartungen an eine KI-basierte Analyseanwendung sollte man die bekannten Vorbedingungen und Einschränkungen genau studieren. Diese sind meist in der Anwendungsdokumentation aufgeführt.

Jede Überwachungsanwendung ist einzigartig, und die Leistung der Anwendung muss direkt am jeweiligen Standort untersucht werden. Wird die erwartete oder vorausberechnete Qualität nicht erreicht, darf man die Untersuchung nicht nur auf die Anwendung selbst beschränken. Alle Untersuchungen müssen auf ganzheitlicher Ebene angestellt werden, denn die Leistung einer Analyseanwendung hängt von sehr vielen Faktoren ab. Die meisten von ihnen lassen sich optimieren, wenn man ihre Auswirkungen kennt. Diese Faktoren sind zum Beispiel die Kamera-Hardware, Videoqualität, Szenendynamik, Beleuchtung sowie Kamerakonfiguration, -position und -ausrichtung.

8.1 Bildnutzbarkeit

Hohe Bildqualität wird oft mit hoher Auflösung und hoher Lichtempfindlichkeit gleichgesetzt. Diese Faktoren sind zweifellos wichtig, aber wie *nützlich* ein Bild oder Video tatsächlich ist, wird durch weitere Faktoren bestimmt. Der hochwertigste Videostream aus der teuersten Überwachungskamera ist nutzlos, wenn eine Nachtszene nicht ausreichend beleuchtet ist, die Kamera abgelenkt oder die Systemverbindung unterbrochen wurde.

Die Platzierung der Kamera muss vor der Installation genau überlegt werden. Damit die Videoanalyse wie erwartet funktionieren kann, muss die Kamera so montiert sein, dass sie eine klare und unbehinderte Sicht auf die beabsichtigte Szene bietet.

Die Nutzbarkeit der Bilder kann auch vom Einsatzzweck abhängen. Videos, die für das menschliche Auge gut aussehen, müssen nicht unbedingt die optimale Qualität für eine Videoanalyseanwendung haben. Tatsächlich werden viele Bildverarbeitungsverfahren zur Verbesserung der Videoansicht für den menschlichen Betrachter nicht empfohlen, wenn eine Videoanalyse erfolgen soll. Dies sind beispielsweise Verfahren zur Rauschunterdrückung, Verfahren für einen großen Dynamikbereich oder Algorithmen zur Belichtungskorrektur.

Viele aktuelle Videokameras verfügen über integriertes Infrarotlicht, damit sie auch bei völliger Dunkelheit funktionieren. Das ist gut, weil die Kameras so auch an Orten mit schwierigen Lichtbedingungen installiert werden können, ohne dass man zusätzliche Beleuchtung vorsehen muss. Werden jedoch starker Regen oder Schneefall erwartet, sollte man sich nicht auf Licht von der Kamera oder aus ihrer direkten Umgebung verlassen. Zu starkes Licht könnte an Regentropfen oder Schneeflocken direkt zur Kamera zurück reflektiert werden, so dass keine Analyse mehr möglich ist. Bei Umgebungslicht stehen die Chancen auf Analyseergebnisse sogar bei schwierigen Witterungsbedingungen besser.

8.2 Erfassungsreichweite

Die maximale Erfassungsreichweite einer KI-basierten Analyseanwendung ist nur schwer abschätzbar. Der Wert in Metern oder Fuß im Datenblatt ist nie ganz aussagekräftig. Bildqualität, Szenencharakteristik, Wetterbedingungen und Objekteigenschaften wie Farbe und Helligkeit beeinflussen die Erfassungsreichweite deutlich. So kann ein helles Objekt vor dunklem Hintergrund an einem sonnigen Tag viel leichter visuell erkannt werden als ein dunkles Objekt an einem regnerischen Tag.

Die Erfassungsreichweite hängt außerdem von der Geschwindigkeit des zu erkennenden Objekts ab. Um exakte Ergebnisse zu erhalten, muss eine Videoanalyseanwendung das Objekt ausreichend lange Zeit „sehen können". Diese Zeit hängt von der Rechenleistung (Bildrate) der Plattform ab: Je geringer die Rechenleistung, desto länger muss das Objekt sichtbar sein, um erkannt zu werden. Ist die Verschlusszeit der Kamera nicht gut auf die Objektgeschwindigkeit abgestimmt, kann Bewegungsunschärfe im Bild auch die Erkennungsgenauigkeit beeinträchtigen.

Schnelle Objekte werden leichter übersehen, wenn sie die Kamera nah passieren. Eine rennende Person, die sich weit von der Kamera entfernt befindet, kann z. B. gut erkannt werden, während eine Person, die sehr nahe an der Kamera mit der gleichen Geschwindigkeit vorüber rennt, so schnell in das Sichtfeld hinein- und wieder herauslaufen kann, dass kein Alarm ausgelöst wird.

Bei Analysefunktionen auf der Grundlage der Bewegungserkennung stellen Objekte, die sich direkt auf die Kamera zu oder von ihr weg bewegen, eine weitere Herausforderung dar. Die Erkennung ist bei langsamen Objekten, die das angezeigte Bild nur geringfügig verändern, im Vergleich zu Bewegungen durch die Szene besonders schwierig.

Höhere Kameraauflösung bedeutet in der Regel keine größere Erfassungsreichweite. Die erforderlichen Verarbeitungsressourcen für die Erstellung eines Algorithmus für maschinelles Lernen sind proportional zur Menge der Eingabedaten. Die erforderliche Rechenleistung für die Analyse einer 4K-Kamera in voller Auflösung ist mindestens vier Mal höher als bei einer 1080p-Kamera. Sehr oft werden KI-basierte Anwendungen wegen der Einschränkungen bei der Rechenleistung der Kamera bei geringerer Auflösung ausgeführt, als es die Kamera oder der Stream ermöglichen würde.

8.3 Alarm- und Aufzeichnungs-Setup

Wegen der diversen Filterstufen generiert die Objektanalytik nur wenige Falschalarme. Aber die Objektanalytik funktioniert nur ordnungsgemäß, wenn alle aufgeführten Vorbedingungen erfüllt sind. Andernfalls könnte sie sogar wichtige Ereignisse verpassen.

Wenn man nicht absolut sicher ist, dass sämtliche Bedingungen jederzeit erfüllt werden, wird deshalb eine konservative Vorgehensweise empfohlen, bei der man das System so einrichtet, dass Alarme nicht nur bei einer bestimmten Objektklassifizierung ausgelöst werden. Dies erzeugt mehr Falschalarme, reduziert aber auch die Gefahr, etwas Wichtiges zu verpassen. Werden Alarme oder Auslöser direkt an eine Alarmzentrale weitergeleitet, kann jeder Falschalarm sehr teuer werden. Es besteht die klare Notwendigkeit für eine zuverlässige Objektklassifizierung, die unerwünschte Alarme herausfiltert. Doch die Aufzeichnungslösung kann und sollte trotzdem so eingerichtet werden, dass sie sich nicht nur auf die Objektklassifizierung verlässt. Bei einem verpassten echten Alarm kann man dann anhand der Aufzeichnung den Grund ermitteln, warum ein Alarm nicht erkannt wurde, und die Installation und Konfiguration entsprechend verbessern.

Wird die Objektklassifizierung während der Suche nach einem Vorfall auf dem Server ausgeführt, sollte das System für kontinuierliche Aufzeichnung konfiguriert werden, so dass die ungefilterte Originalaufzeichnung erhalten bleibt. Kontinuierliches Aufzeichnen verbraucht viel Speicherplatz, aber moderne Komprimierungsalgorithmen wie Zipstream können dies teilweise ausgleichen.

8.4 Wartung

Eine Überwachungsinstallation muss regelmäßig gewartet werden. Neben der Betrachtung des Videos durch die VMS-Schnittstelle werden auch physische Inspektionen empfohlen, um alles zu entfernen, was das Sichtfeld stören oder blockieren könnte. Das ist zwar auch bei Standardinstallationen ratsam, die lediglich aufzeichnen, doch bei Einsatz von Analysefunktionen ist es unverzichtbar.

Im Kontext einer allgemeinen Bewegungserkennung in Videos könnte ein typisches Hindernis wie ein im Wind schwingendes Spinnennetz die Anzahl der Alarme und dadurch den Speicherbedarf unnötig erhöhen. Bei der Objektanalyse würde das Netz einen Ausschlussbereich im Erkennungsbereich schaffen. Die einzelnen Fäden würden Objekte verdecken und die Fähigkeit zur Erfassung und Klassifizierung drastisch einschränken.

Spinnweben können das Sichtfeld einer Überwachungskamera stören.

Schmutz auf der Frontscheibe oder Abdeckung verursachen tagsüber eher keine Probleme. Aber bei schlechtem oder seitlich auf die Abdeckung auftreffendem Licht beispielsweise von Autoscheinwerfern können sie unerwartete Reflexionen hervorrufen, die die Erfassungsgenauigkeit reduzieren.

Eine szenenbezogene Systempflege ist genauso wichtig wie die Kamerawartung. Während der Lebensdauer einer Kamera kann im von ihr überwachten Bereich viel passieren. Ein einfacher Vorher-Nachher-Vergleich der Bilder deckt mögliche Probleme auf. Wie sah die Szene bei Einrichtung der Kamera aus und was zeigt sie heute? Muss der Erfassungsbereich angepasst werden? Muss das Sichtfeld der Kamera angepasst oder die Kamera an einen anderen Ort versetzt werden?

9 Datenschutz und Unversehrtheit der Person

Die Arbeit in der Sicherheits- und Überwachungsbranche ist immer mit einer Abwägung des Rechtes auf Datenschutz und persönliche Integrität der Person gegen den Schutz vor Verbrechen oder die Schaffung von Möglichkeiten für forensische Untersuchungen verbunden. Dies erfordert sorgfältige ethische Überlegungen am Installationsort für den jeweiligen Einsatzzweck sowie ein Verständnis und die Anwendung der örtlichen Rechtslage. Außerdem muss die Lösung oft noch weitere Aufgaben erfüllen und zum Beispiel die Cybersicherheit gewährleisten und unbefugte Zugriffe auf Videomaterial verhindern. Edge-basierte Analytik und die Erzeugung von Metadaten zu statistischen Zwecken könnte jedoch auch den Datenschutz stärken, wenn nur anonymisierte Daten zur späteren Verarbeitung übertragen werden.

Mit der Zunahme der automatisierten Analyse in Überwachungssystemen müssen einige neue Aspekte berücksichtigt werden. Weil die Analyseanwendungen die Gefahr von Fehlalarmen bergen, muss immer ein erfahrener Bediener/eine Wachperson an der Entscheidung beteiligt sein. Dies wird oft als „Mensch in der Schleife" bezeichnet. Außerdem ist es wichtig zu erkennen, dass die menschliche Entscheidung davon beeinflusst werden kann, wie der Alarm generiert und präsentiert wird. Ohne ordnungsgemäße Schulung und Wissen um die Funktionalität der Analyselösung könnte er die falschen Schlüsse ziehen.

Ein weiteres Problem könnte die Entwicklungsweise der Algorithmen für Deep Learning sein. Manche Nutzungsfälle erfordern große Sorgfalt bei der Anwendung der Technologie. Die Qualität dieser Algorithmen ist elementar mit den Datensätzen verknüpft, also mit den Videos und Bildern, mit denen der Algorithmus trainiert wird. Tests haben gezeigt, dass bestimmte KI-Systeme bei der Erkennung beginnen können, ethnisch oder geschlechtsspezifisch zu diskriminieren, wenn das Material nicht sorgfältig ausgewählt wird. Dies hat eine offene Diskussion angeregt und zu rechtlichen Beschränkungen und Aktivitäten angeregt, die dafür sorgen sollen, dass diese Aspekte bei der Entwicklung der Systeme berücksichtigt werden.

Da die KI immer häufiger im Überwachungsbereich eingesetzt wird, müssen die Vorteile der betrieblichen Effizienz und neuer möglicher Anwendungsfälle immer von einer sorgfältigen Überlegung begleitet werden, wann und wo sich ihr Einsatz wirklich lohnt.

10 Anhang

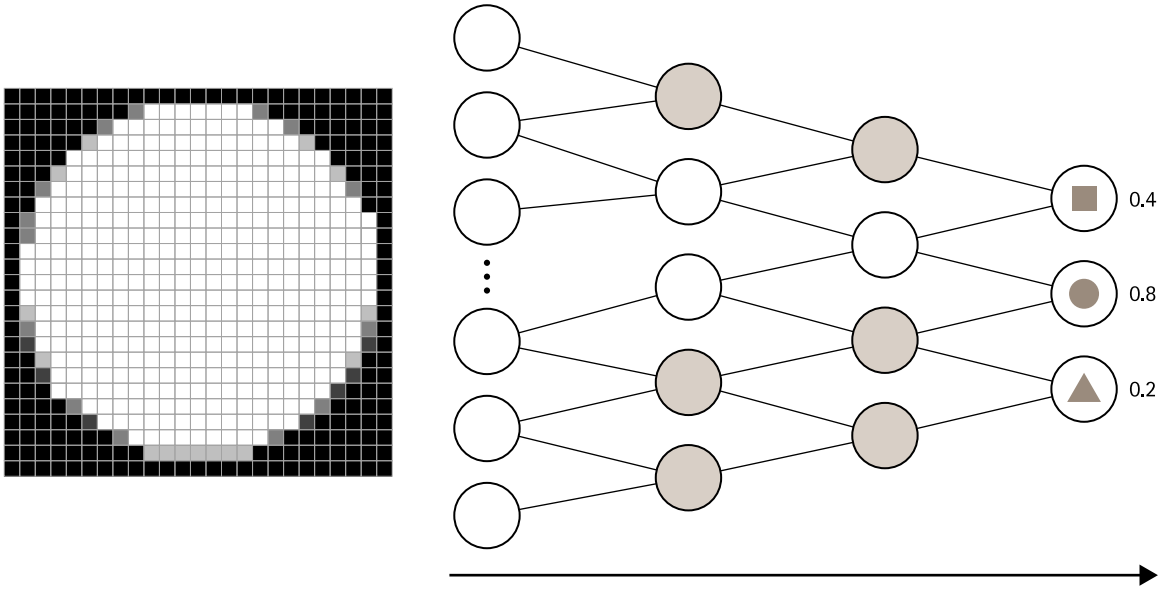
Dieser Anhang liefert Hintergrundinformationen zu künstlichen neuronalen Netzwerken, die eine Grundlage für Deep Learning bilden.

10.1 Neuronale Netzwerke

Neuronale Netzwerke sind eine Familie von Algorithmen, die auf ähnliche Weise wie das menschliche Gehirn Beziehungen in Datensätzen erkennen können. Ein neuronales Netzwerk besteht aus hierarchischen, miteinander verbundenen Schichten, den so genannten Knoten oder Neuronen. Informationen werden entlang den Verknüpfungsstellen von der Eingabeschicht durch das Netzwerk bis zur Ausgabeschicht weitergereicht.

Damit neuronale Netzwerke funktionieren können, geht man davon aus, dass Eingabe-Probedaten auf einen endlichen Satz von Merkmalen reduziert werden können, die eine ausreichende Darstellung der Eingabedaten liefern. Diese Merkmale können daraufhin kombiniert werden, um die Eingabedaten zu klassifizieren, also beispielsweise den Inhalt eines Bilds zu beschreiben.

Die folgende Abbildung zeigt ein Beispiel für ein neuronales Netzwerk, das erkennt, zu welcher Klasse das Eingangsbild gehört. Jedes Pixel im Bild entspricht einem Eingangsknoten. Alle Eingangsknoten sind mit den Knoten in der ersten Schicht verknüpft. Diese erstellen Ausgangswerte, die als Eingangswerte an die zweite Schicht weitergeleitet werden, und so weiter. In jeder Schicht werden außerdem Gewichtung, Bias-Werte und Aktivierungsfunktionen berücksichtigt.



Beispiel für ein Eingangsbild (links) und ein neuronales Netzwerk (rechts). Bis zur Ausgabeschicht hat das Netzwerk Wahrscheinlichkeiten für jede mögliche Kategorie (Quadrat, Kreis oder Dreieck) ermittelt. Die Kategorie mit dem höchsten Wahrscheinlichkeitswert ist die wahrscheinlichste Form des Eingangsbilds.

Diesen Vorgang bezeichnet man als *Vorwärtspropagation.* Weicht das Ergebnis der Vorwärtspropagation vom Zielwert ab, werden die Netzwerkparameter durch *Backpropagation* leicht verändert. Durch dieses iterative Training wird die Performance des Netzwerks schrittweise verbessert.

Nach dem Deployment erinnert sich ein neuronales Netzwerk normalerweise nicht an die früheren Durchläufe. Es kann sich also im Laufe der Zeit nicht mehr verbessern und kann nur die Arten von Objekten erkennen, bzw. die Arten von Aufgaben ausführen, für die es trainiert wurde.

10.2 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) sind ein Subtyp künstlicher neuronaler Netze, die sich als besonders geeignet für Computer-Vision-Aufgaben erwiesen haben und im Mittelpunkt des rasanten Fortschritts des Deep Learning stehen. Bei Computer Vision wird das Netzwerk darauf trainiert, automatisch nach bestimmten Bildmerkmalen wie Begrenzungen, Ecken und Farbunterschieden zu suchen und dadurch Formen in einem Bild zu erkennen.

Dies geschieht hauptsächlich durch eine mathematische Operation namens *Convolution* (Faltung). Diese läuft sehr effizient ab, weil die Ausgabe jedes einzelnen Knotens nur von einer begrenzten Umgebung der Eingabedaten und nicht von der gesamten Eingabedatenmenge abhängig ist. Anders ausgedrückt: In einem CNN ist nicht jeder Knoten mit jedem Knoten in der vorigen Schicht, sondern nur mit einer kleinen Untergruppe davon verbunden. Die Konvolutionen werden durch weitere Operationen zur Reduzierung der Datenmenge ergänzt, wobei die nützlichsten Informationen bewahrt werden. Wie in einem regulären künstlichen neuronalen Netz werden die Daten immer abstrakter, je weiter sie im Netz wandern.

Während der Trainingsphase lernt das CNN, wie es die Schichten am besten anwendet. Das heißt, wie die Konvolutionen die Merkmale aus der vorigen Schicht kombinieren müssen, damit die Ausgabe des Netzwerks möglichst genau den Anmerkungen der Trainingsdaten entspricht. Bei der Schlussfolgerung wendet das trainierte Convolutional Neural Network daraufhin die Schichten der Konvolutionen, die es gelernt hat, nacheinander an.

T10163986/DE/M2.2/2103

Informationen zu Axis Communications

Axis ermöglicht eine smarte und sichere Welt durch die Entwicklung von Netzwerk-Lösungen. Diese bieten Erkenntnisse, um die Sicherheit und Geschäftsmethoden zu verbessern. Als Technologieführer im Bereich Netzwerk-Video bietet Axis Produkte und Dienstleistungen für die Videoüberwachung/-analyse und Zutrittskontrolle sowie Sprechanlagen und Audiosysteme. Das 1984 gegründete schwedische Unternehmen beschäftigt mehr als 3.800 engagierte Mitarbeiter in über 50 Ländern. Gemeinsam mit seinen Partnern auf der ganzen Welt bietet das Unternehmen kundenspezifische Lösungen an.

Weitere Informationen über Axis finden Sie unter *axis.com*.

AXIS COMMUNICATIONS