VON DER KI-THEORIE ZU ECHTEM WIRTSCHAFTLICHEN NUTZEN

Infrastruktur-Überlegungen für IT-Verantwortliche

INHALT

Jetzt direkt zum Abschnitt gelangen, der Ihnen die Vorteile der neuesten Generation der skalierbaren Intel® Xeon® Prozessoren und Intel® Deep Learning Boost präsentiert

1. DEEP LEARNING WIRD ERWACHSEN … UND ZWAR RASEND SCHNELL

Bis 2020 wird Deep Learning ein vollkommen anderes Reifestadium erreicht haben. Bereitstellung und Einführung wird nicht mehr auf Versuche beschränkt bleiben. Vielmehr wird Deep Learning quer über die meisten Forschungsgebiete und Branchen hinweg zu einem wesentlichen Bestandteil des täglichen Geschäfts werden.

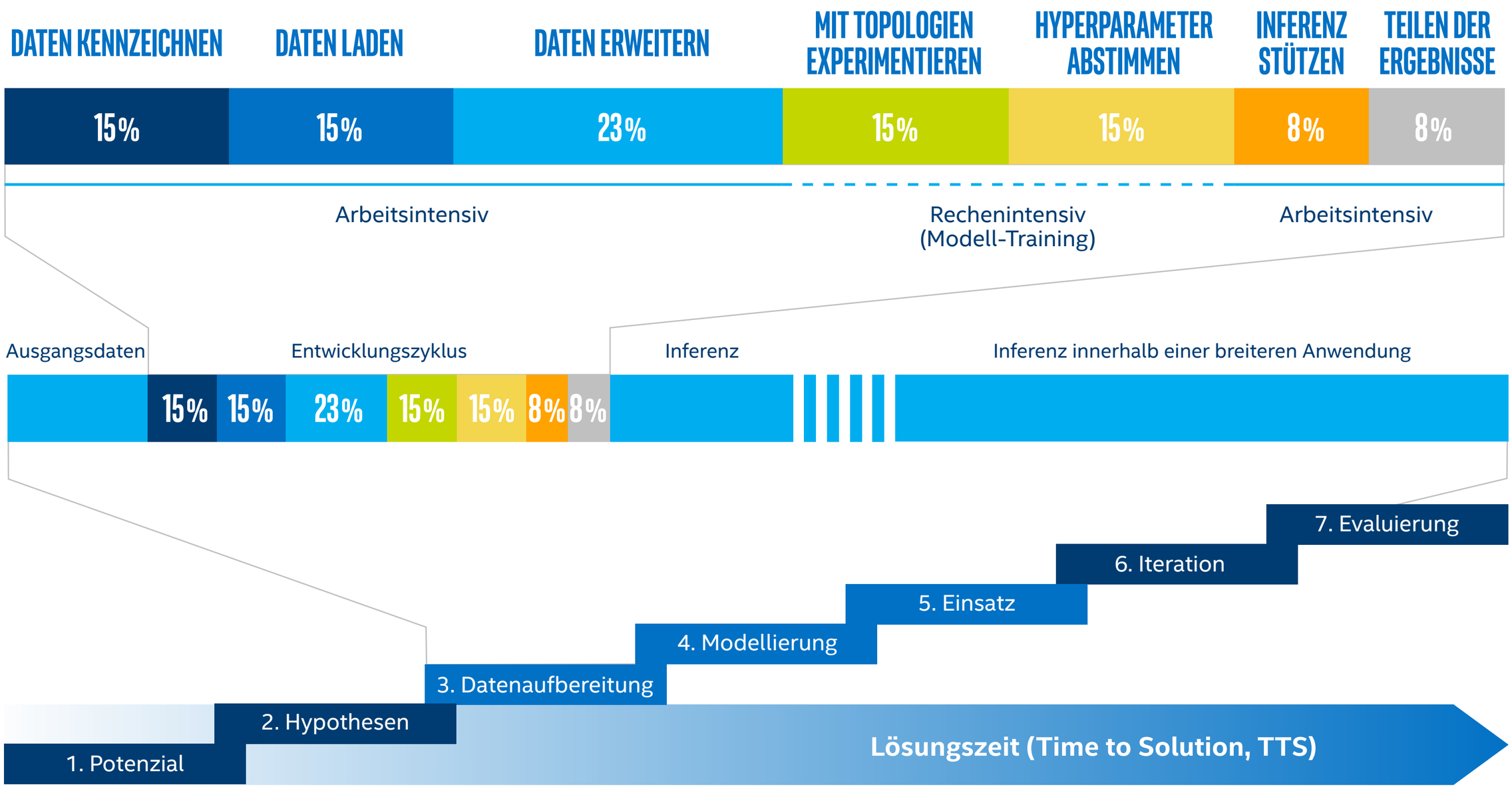
Wieso? Weil die für Deep-Learning-Aufgaben eingesetzte Hard- und Software bei Geschwindigkeit und Genauigkeit Fortschritte gemacht hat. Dadurch ist sie auch praktikabel und kosteneffizient geworden. Ein Großteil dieses Mehrwerts wird durch Deep-Learning-Inferenz geschaffen – das bedeutet, dass ein Modell verwendet wird, das Erkenntnisse aus ihm zuvor unbekannten Daten ableitet. Solche Modelle können in der Cloud oder im Rechenzentrum eingesetzt werden. Künftig werden wir sie jedoch immer öfter auf Endgeräten wie Kameras oder Smartphones sehen.

Laut Prognosen von Intel wird sich das Verhältnis zwischen Inferenz- und Trainingszyklen von 1:1 in den frühen Tagen des Deep Learning auf deutlich über 5:1 bis 2020 verschieben.[1] Intel nennt diese Verschiebung „Inferenz in großem Maßstab“. Da Inferenz auch beinahe 80 Prozent der Workflows von künstlicher Intelligenz (KI) ausmacht (Abbildung 1, Seite 3), folgt daraus, dass der Weg zu echter KI-Bereitschaft mit der Auswahl von Hardware-Architekturen beginnt, die für diese Aufgabe gut geeignet sind.

Da jedoch der Bereich der KI immer komplexer wird, kann eine Universallösung nicht alle spezifischen Bedingungen einer jeder Umgebung im KI-Spektrum abdecken. In diesem Zusammenhang gehören Verfügbarkeit, Benutzerfreundlichkeit und Betriebskosten zu den wichtigen Hardware-Anforderungen. Welche Art von Infrastruktur kommt aktuell bei Edge-Geräten, Workstations und Servern zum Einsatz? Besteht die Bereitschaft, sich mit der Komplexität unterschiedlicher Architekturen zu beschäftigen?

Die Auseinandersetzung mit diesen Herausforderungen ist das Thema dieses Papers.

ABBILDUNG 1: EIN TYPISCHER KI-WORKFLOW



2. ERMITTLUNG DER KI-BEREITSCHAFT

In welchem Stadium der KI-Bereitschaft sich ein Unternehmen befindet, ist entscheidend für die Priorisierung seiner Maßnahmen, und ebnet auch den Weg vom Experimentieren mit KI hin zu Praxiseinsätzen. Die KI-Bereitschaft von Unternehmen lässt sich in drei Kategorien einteilen: grundsätzliche Bereitschaft, Einsatzbereitschaft und Transformationsbereitschaft. Der Schritt in die nächste Phase oder dauerhafter Erfolg erfordern die passenden Elemente hinsichtlich Fähigkeiten und Ressourcen, Infrastruktur und Technologie sowie Prozessen und Modellen.

Zu den entscheidenden Merkmalen von Unternehmen mit Einsatz- und Transformationsbereitschaft für KI zählen – in jeweils unterschiedlichem Maße – ihre Fähigkeit, durch KI mit Inferenz in großem Maßstab eine bessere Entscheidungsfindung oder Automatisierung von Geschäftsvorgängen/Reaktionen zu ermöglichen.

In der Phase der grundsätzlichen Bereitschaft sollten Unternehmen Entwicklung und Einsatz von Proofs of Concept (PoCs) priorisieren, um die zur Skalierung der KI notwendige Infrastruktur, Kompetenz und Akzeptanz der Führungskräfte auf- und auszubauen.

Diese Themen werden in unseren Whitepapers In 5 Schritten zum KI-Proof-of-Concept und Das KI-Bereitschaftsmodell eingehend behandelt.

Bereitschaftsstufe	Top-Priorität der IT	Bereitschaftsindikatoren
Grundsätzliche Bereitschaft	Sobald eine Datenstrategie und vorgesehene Anwendungsfälle identifiziert wurden, ist für eine Infrastruktur und für Schnittstellen zu sorgen, die für einen KI-PoC angemessen sind.	• KI beginnt bei den Daten. Es muss sichergestellt werden, dass wichtige Datenquellen verfügbar und zugänglich sind. • Kapazitäten und Fähigkeiten von Rechenzentren, um einen PoC durch die stark skalierbare Verarbeitung zu unterstützen, die für KI benötigt wird. • Dank einer niedrigen Einstiegsschwelle und Pay-Per-Use-Services könnte erkundet werden, inwiefern sich Cloud-basierte Services für das Testen von Anwendungsfällen nutzen lassen. • Überlegungen, wie sich Open-Source- und kommerzielle KI-Softwarepakete in andere Tools für Datenverwaltung und Visualisierung integrieren lassen.
Einsatzbereitschaft	Geeignete Management- und Governance-Mechanismen für eine nachhaltige Entwicklung von KI-Lösungen gehören eingerichtet.	• Bewährte Modelle wie DevOps werden eingesetzt, um auf schnell wechselnde Anforderungen zu reagieren. • Falls erforderlich, werden für einen spezifischen Anwendungsfall notwendige Fähigkeiten intern aufgebaut, um über einen ersten PoC hinauskommen zu können. • Die Sicherheit von Daten, Infrastruktur und Algorithmen wird priorisiert, um Risiken aufgrund fehlerhafter Dateneingaben, Manipulationen am Modell und unerlaubter Zugriffe auf die gewonnenen Erkenntnisse zu verringern.
Transformationsbereitschaft	Die Fähigkeit des Unternehmens, KI optimal zu nutzen, sollte priorisiert werden. Unterstützt sie eine bessere Entscheidungsfindung auf Führungsebene oder automatisiert sie Geschäftsvorgänge bzw. fördert sie automatisierte Reaktionen?	• Konsequente Arbeit interner Stakeholder am Aufbau einer Organisationsstruktur, die KI-getriebene Möglichkeiten zur Verbesserung von Abläufen und Interaktionen mit Kunden identifiziert. • Es liegt ein genau berechnetes Geschäftsszenario dafür vor, wie Erfolg mit KI aussieht. • Ein Fokus auf die Akzeptanz von KI stellt sicher, dass die Lösung an die geschäftlichen Bedürfnisse angepasst wird – bis hin zu den täglichen Aktivitäten des Personals mit Kundenkontakt und den Betroffenen.

3. ENTWICKLUNG UND EINSATZ VON RICHTLINIEN FÜR DATA GOVERNANCE UND SICHERHEIT

Sicherheit lässt sich im Kontext von KI aus zwei verschiedenen Perspektiven betrachten. Erstens ist es wichtig, die KI selbst in der Form von Algorithmen, Parametern und Daten zu sichern. Zweitens besitzt KI großes Potenzial, das sich für das Auffinden raffinierter Exploits nutzen lässt.

Die Wechselbeziehungen zwischen KI, Sicherheit und Governance sind komplex und vielschichtig. In den frühen Unternehmensphasen unterscheiden sich die Governance-Fragen nicht von denen anderer datenbasierter IT-Projekte: Kann das Projekt Ergebnisse liefern, ist der Schutz der Kundendaten gewährleistet usw.? Mit zunehmender Nutzung der KI ergeben sich zusätzliche Auswirkungen: Wie viel menschliches Eingreifen ist beispielsweise bei der vorausschauenden Planung und Instandhaltung (gegebenenfalls) für Einkaufsentscheidungen erforderlich?

Wo von einer KI getriebene Entscheidungen Auswirkungen auf das Leben von Menschen haben, besteht zudem das Risiko, dem Ruf des Unternehmens zu schaden, wenn die KI-Ergebnisse ungenau oder verfälscht sind. Diese Gefahr ist weit größer als das aktuelle Reputationsrisiko durch Verletzung des Datenschutzes im Zusammenhang mit Kundendaten.

Solche Bedrohungen könnten in Form von „Modell-Vergiftungen“ auftreten, wenn ein Modell durch abweichende Eingaben oder durch den Einbau von Hintertüren in ungenutzte Parameter verfälscht wird. Hardware-basierte Trusted Execution Environments (TEEs) sollten in Erwägung gezogen werden. Dadurch könnten vertrauenswürdige Modelle zwischen Endpunkt und Aggregator gebaut werden, wo Aktualisierungen geschützt sind. Das würde das Risiko von Modell-Vergiftungen minimieren.

4. INFRASTRUKTURSTRATEGIEN FÜR DEN UMSTIEG AUF DEEP-LEARNING-INFERENZ IN GROSSEM MASSTAB

KI ist eine komplexe Mischung von Aufgaben: Rohdaten werden gebrauchsfertig aufbereitet, Modelle erstellt, gesichert und feinabgestimmt. Die Lösungen werden in großem Maßstab in der realen Welt eingesetzt, wo sie ständig optimiert werden und außerhalb der relativen Sicherheit eines On-Premise-Rechenzentrums laufen – oft mit starker Einschränkung von Serverleistung und Platzangebot.

Das bedeutet, dass der Entwurf einer On-Premise- und/oder Hybrid-Cloud-Lösung für KI einen gänzlich neuen Ansatz erfordert. Dazu gehört auch die Einrichtung flexibler Rechenzentren, die in der Lage sind, umfangreiche On-Demand-Rechenressourcen zu bündeln. KI erfordert auch Speicherressourcen, Konnektivität und – möglicherweise – Netzwerke, die dazu in der Lage sind, Daten bei hoher Geschwindigkeit mit minimaler Latenz zu übertragen.

Weitere Schwierigkeiten ergeben sich aus der Tatsache, dass KI keine Universallösung darstellt. Die Optionen der IT-Verantwortlichen, die nach den richtigen Infrastrukturstrategien für ihr Unternehmen suchen, lassen sich in vier Kategorien unterteilen (diese werden in unserem Whitepaper Infrastruktur-Strategien für individuelle KI-Lösungen näher untersucht).

Umfunktionierung vorhandener Hardware

- **Was:** Unternehmen, die noch am Anfang ihrer KI-Aktivitäten stehen, nutzen oft „freie Zyklen“ in ihren Rechenzentren, um KI-Workloads auszuführen. Oder sie entwickeln Lösungen, die auf einem einzelnen „freien“ Server bzw. Workstation-Node oder einem kleinen Rechnerverbund basieren.
- **Die Vorteile:** Die Nutzung vorhandener Hardware-Ressourcen ermöglicht es der Forschung, sich auf eine eng verbundene Umgebung zu konzentrieren, und sie verstärkt den Nutzen der KI bei der „Verbesserung“ bestehender Fähigkeiten.
- **Die Nachteile:** Eine Integration in umfassendere Lösungen ist nicht immer einfach möglich. Wenn Hardware nicht auf die Anforderungen abgestimmt wurde, können Fixkosten dafür entstehen, dass weniger geeignete Ressourcen anders genutzt oder umverteilt werden.

Aufbau einer breiteren Plattform

- **Was:** Unternehmen mit mehr Erfahrung im Bereich KI wünschen sich möglicherweise eine umfassendere Infrastrukturlösung, die allgemeinere KI-Workloads unterstützt.
- **Die Vorteile:** Ein plattformbasierter Ansatz bietet einen einzigen Konfigurationspunkt und ein einheitliches Bereitstellungsziel. Das lässt sich auch organisatorisch einfacher verwalten und setzt den Schwerpunkt auf den Aufbau von Fachwissen.
- **Die Nachteile:** Der Aufbau könnte anfangs als komplexer angesehen werden. Er erfordert schwer verfügbares firmeninternes Know-how und erhöht das Risiko, falls sich die Architektur als zu klein oder zu groß für den tatsächlichen Bedarf erweist.

Kauf einer Einzellösung

- **Was:** Eine maßgeschneiderte Lösung, die dafür angeschafft wurde, um einen genau definierten Anwendungsfall zu erfüllen.
- **Die Vorteile:** Sorgt potenziell für schnelleren Einsatz, Effizienz und Performance, da die Lösung in Hinblick auf einen vorgegebenen Anwendungsfall entwickelt wurde.
- **Die Nachteile:** Ein Einzellösung kann zu mehreren KI-„Silos“ führen, die parallel verwaltet werden müssen. Insgesamt kann sich das als teurer erweisen.

Outsourcing der Lösungsbereitstellung

- **Was:** Unternehmen in verschiedenen Stadien auf dem Weg zur KI können Ressourcen von Drittanbietern (einschließlich Cloud-basierter Optionen) nutzen, um entweder eine Komplettlösung bereitzustellen oder mit vorhandenen Ressourcen zu arbeiten.
- **Die Vorteile:** Funktionalitäten sind „von der Stange“ verfügbar, wodurch Probleme bei der Bereitstellung und Konfiguration minimiert werden. Externe Dienstleistungen eignen sich zur Erweiterung und Erprobung neuer Lösungen, bevor sie intern eingeführt werden.
- **Die Nachteile:** Mehrkosten und Ineffizienz durch die Koordination von Dienstleistern. Die entstehende Infrastrukturarchitektur kann zu Datenengpässen führen, je nachdem, woher die Daten stammen. Das Unternehmen muss etwa unter Umständen Daten aus seinen internen Systemen in die Cloud hochladen.

JENSEITS DES RECHENZENTRUMS: WIE KI AN DEN NETZWERKRAND GEBRACHT WIRD

Der Großteil von KI geschieht aktuell in Rechenzentren oder in der Cloud. Da sich Milliarden von Geräten mit dem Internet verbinden und der Bedarf an Echtzeit-Intelligenz wächst, wird mehr KI-Inferenz an den Netzwerkrand verschoben, damit Daten nicht in die Cloud übertragen werden müssen.

Eine der sichersten Methoden, KI an den Netzwerkrand zu verschieben, ist das sogenannte Federated Learning. Dieser Prozess ermöglicht es Endgeräten, gemeinschaftlich ein gemeinsames Vorhersagemodell zu erlernen. Dabei bleiben alle Trainingsdaten auf den Geräten. So wird die Möglichkeit, Modelle zu verbessern, davon entkoppelt, die Daten in der Cloud speichern zu müssen. Dadurch wird es auch möglich, die Geräte für das Modell-Training zu nutzen. Das Gerät lädt sich das neueste Modell herunter und verbessert es, indem es von den Daten auf dem Gerät lernt und die Änderungen zu einer kleinen, gezielten Aktualisierung zusammenfasst. Nur diese Aktualisierung wird mittels verschlüsselter Kommunikation an die Cloud geschickt. Dort wird sie sofort mit anderen Nutzer-Aktualisierungen gemittelt, um das gemeinsame Modell zu verbessern. Alle Trainingsdaten verbleiben auf dem Endgerät und keine individuellen Aktualisierungen werden in der Cloud gespeichert.

Intel bietet Unternehmen Hard- und Software-Tools für den Einsatz von KI auf Endgeräten. Dazu gehören:

- Das Intel® Distribution for OpenVINO™ Toolkit wurde für neuronale Netzwerke von Videos entwickelt, die auf unterschiedlicher Hardware von Intel zum Einsatz kommen - vom Rechenzentrum bis zu Geräten am Netzwerkrand.

MEHR ERFAHREN >

- Intel® Movidius™ Vision Processing Units (VPU), die die Grenzen dessen erweitern, was mit KI am Netzwerkrand möglich ist – durch extrem stromsparende Deep-Neural-Network-Inferenzen (DNN) auf dem Gerät.

MEHR ERFAHREN >

Aufgrund ihrer Art verlangen viele KI-Anwendungsfälle von den Systemen, dass sie in Echtzeit Schlussfolgerungen ausführen - und nicht offline oder im Batch-Modus. Außerdem müssen Modelle möglicherweise im Laufe der Zeit neu trainiert und aktualisiert werden. Jedoch können sowohl Speicher, Prozessor als auch Datenübermittlung zu Engpässen führen, die die Auslastung senken und die Kosten erhöhen.

Intel bietet Hardware- und Speicher-Lösungen für Unternehmen, die vor diesen Herausforderungen stehen:

- Aktuell laufen weltweit bereits viele Inferenz-Workloads in Rechenzentren auf **skalierbaren Intel® Xeon® Prozessoren**. Diese Prozessoren wurden speziell dafür weiterentwickelt, neben Rechenzentrums- und Cloud-Anwendungen, die bereits darauf laufen, jetzt auch High-Performance-KI-Anwendungen zu unterstützen.
- Dank **Intel® Deep Learning Boost (Intel® DL Boost)** bietet die 2. Generation der skalierbaren Intel® Xeon® Prozessoren eine höhere KI-Beschleunigung. Intel® DL Boost enthält eingebettete Befehlssätze (Vector Neural Network Instructions, VNNI), die komplexe Berechnungen beschleunigen, wie sie für Convolutional Neural Networks (CNN) und andere Deep Neural Networks (DNN) typisch sind. Das Ergebnis ist eine effizientere Inferenzbeschleunigung für Deep-Learning-Anwendungsfälle wie Bildklassifizierung, Spracherkennung, Sprachübersetzung und Objekterkennung. Bei Tests erzielte die 2. Generation der Intel® Xeon® Platinum 8280 Prozessoren mit Intel® DL Boost einen 14-mal so hohen Inferenzdurchsatz[2] wie die Dual-Sockel Intel® Xeon® Platinum 8180 Prozessoren. Der Intel® Xeon® Platinum 9282 Prozessor brachte sogar noch eine weitere Steigerung: Er erzielte bei der Bildklassifizierung eine herausragende Inferenzleistung, die 30-mal so hoch war[3] wie die des oben genannten älteren Modells.
- Die 2. Generation der skalierbaren Intel® Xeon® Prozessoren unterstützt außerdem persistenten Intel® Optane™ DC Speicher. Dadurch steht unmittelbar bei der CPU mehr Speicher zur Verfügung, sodass Daten auch bei Stromausfällen und Systemwartungen erhalten bleiben. Dieser Speicher unterstützt auch kürzere Startzeiten, da er nicht-flüchtig ist. Unternehmen können dadurch ihre Infrastruktur patchen, updaten und sichern, während Betriebszeit und Servicebereitstellung maximiert werden. Ungeplante Betriebsunterbrechungen wirken sich außerdem weniger stark auf den Geschäftsbetrieb aus, da die Wiederherstellungszeit verkürzt ist. Niedrige Latenz ermöglicht Unternehmen auch, umfangreichere Arbeitsdaten in In-Memory-Datenspeichern zu behalten. So kann aus wesentlich größeren Datenmengen mehr Nutzen gezogen werden, wodurch wichtige Geschäftsentscheidungen schneller getroffen werden können.
- Mit **Intel® Optane™ SSDs** können Unternehmen Speicherengpässe vermeiden. Durch sie können in Rechenzentren große Datenmengen kostengünstiger verarbeitet und Anwendungen beschleunigt werden. Der Einsatz größerer Speicher-Pools ermöglicht zudem das Sammeln unternehmensrelevanter Erkenntnisse. Intel ® Optane™ Technologie bietet auf diese Weise sowohl für das Deep-Learning-Training als auch für die Inferenz einen Mehrwert. Sie optimiert auch das Batch-Training, wodurch sehr viel größere Datenmengen möglich und gleichzeitig die Kapazitäts- und Kostenvorteile von Speicher genutzt werden. Durch beschleunigte Echtzeit-Inferenz wächst das KI-Potenzial erheblich, zudem werden durch die optimierte Batch-Inferenz Analysen effizienter durchgeführt und Erkenntnisse schneller gewonnen.

- Mittels **Intel® Select Solutions für BigDL auf Apache Spark*** werden Hardware-Komponenten zu einer integrierten Plattform zusammengefügt. Durch eine geprüfte, auf Apache Spark basierende Infrastruktur, werden Entwicklung und Einsatz von Deep Learning vereinfacht. Indem Analysen von Daten direkt an dem Ort durchgeführt werden können, an dem sie gespeichert sind, entfällt das Übertragen und Duplizieren von Daten, wodurch KI-Innovation beschleunigt werden. Da KI-Workloads zudem auf der bereits bestehenden Architektur laufen können, die auf Intel® Xeon® Prozessoren basiert, senkt das die TCO, indem die Auslastung erhöht und die IT-Kosten gesenkt werden. Dank umfangreicher Tools und Bibliotheken kann die Wertschöpfung zusätzlich beschleunigt werden.
- **Intel® Select Solutions für das KI-Inferencing** ist eine einsatzfertige Plattform für eine Inferenz mit geringer Latenz und hohem Durchsatz, die direkt auf der CPU und nicht auf einer separaten Beschleunigerkarte durchgeführt wird. Mit Intel® Select Solutions können Sie schnell und einfach Algorithmen für effizientes KI-Inferencing implementieren, und das auf einer Architektur, die aus von Intel geprüften Bausteinen besteht. Intel® Select Solutions nutzt Intel® DL Boost der neuen skalierbaren Intel® Xeon® Prozessoren, um KI-Inferencing zu beschleunigen. Das wird erreicht, indem Inferenzberechnungen mit nur einem Befehl durchgeführt werden, für die zuvor eine ganze Reihe von Befehlen notwendig waren. Mit der Intel® Distribution für OpenVINO™ Toolkits können Sie die Inferenzleistung noch weiter beschleunigen.

Führen Sie anspruchsvolle KI-Inferencing-Workloads auf der 2. Generation der skalierbaren Intel® Xeon® Prozessoren aus – damit vereinfachen Sie Ihre IT-Infrastruktur, sparen Kosten und optimieren die Auslastung.

5. POSITIVE AUSWIRKUNGEN OPTIMIERTER SOFTWARE

Hardware ist nichts ohne die richtige Software, durch die sie Spitzenleistungen erbringt. Jeder KI-Anwendungsfall erfordert eine Software-Architektur, die die besten Tools für die zu erledigende Aufgabe auswählt. Dabei werden nachgelagerte Systeme, Anpassungen, Optimierungen und andere Modifikationen potenziell berücksichtigt. Eine Auswahl davon ist in der nachfolgenden Tabelle zu finden.

Beim Prüfen dieser Optionen ist es wichtig, auf folgende Eigenschaften von Toolkits, Bibliotheken und Frameworks zu achten:

- **Optimiert für bestehende Umgebungen**. Viele Open-Source-Bibliotheken und -Frameworks wie TensorFlow*, MXNet*, PaddlePaddle* und PyTorch* sind für skalierbare Intel® Xeon® Prozessoren optimiert. Intel hat mit Google an TensorFlow*, mit Apache an MXNet* und mit Baidu an PaddlePaddle* und an Caffe* gearbeitet, um die Deep-Learning-Leistung zu verbessern. Bei den skalierbaren Intel® Xeon® Prozessoren wurden entsprechende Software-Optimierungen vorgenommen, weitere Frameworks von Microsoft* und anderen Branchenführern folgen.
- **Abgestimmt darauf, was der Anwendungsfall des Unternehmens erfordert.** IT-Verantwortliche müssen festlegen, welche bestimmten Engpässe oder Probleme durch Software-Optimierungen überwunden werden sollen – egal ob ein PoC durchgeführt oder eine bestehende Lösung skaliert werden sollte.
- **Von den eigenen Mitarbeitern bevorzugt.** Die Datenanalysten und Entwickler des Unternehmens werden mit diesen Tools arbeiten müssen. Wenn sie mit den von der IT ausgewählten Optimierungen nicht vertraut sind, ist es wichtig, dass sie verstehen, wie die Entwicklungszeit reduziert und die Effizienz gesteigert werden kann.

Intel hat mit Google an TensorFlow*, mit Apache an MXNet* und mit Baidu an PaddlePaddle* und an Caffe* gearbeitet, um die Performance von Deep Learning zu verbessern. Dabei wurden Software-Optimierungen für die skalierbaren Intel® Xeon® Prozessoren in Rechenzentren eingesetzt und es werden weitere Frameworks von Microsoft* und anderen Branchengrößen hinzukommen.

TOOLKITS	WIE FUNKTIONIERT DAS?
Intel® Distribution of OpenVINO™ Toolkit	Ein Software-Toolkit, das Computer-Vision-Teams dabei unterstützt, **Entwicklung und Einsatz von auf neuronalen Netzwerken basierenden Anwendungen** auf Gateways und Geräten über mehrere Intel® Plattformen (CPU, GPU, FPGA, VPU) hinweg zu unterstützen. Schnelles Optimieren von bereits trainierten Modellen und ihr Einsatz in einem breiten Spektrum von Intel® Hardware und Beschleunigern, häufig mit erheblichen Leistungssteigerungen durch die Nutzung von Deep-Learning-Frameworks, ohne große Änderungen daran, wie Einsätze aktuell durchgeführt werden.
Intel® Movidius™ Neural Compute SDK (NCSDK)	Ein Softwareentwicklungskit, das **schnelle Prototypenerstellung und den Einsatz von Deep Neural Networks (DNNs)** auf kompatiblen neuronalen Rechnern wie dem Intel® Movidius™ Neural Compute SDK (NCSDK) ermöglicht. Dazu gehören eine Reihe von Software-Tools, um die DNNs zu erstellen, zu trainieren und zu validieren.
BIBLIOTHEKEN	
Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN)	Intel MKL-DNN ist eine leistungssteigernde Open-Source-Bibliothek für die **Beschleunigung von Deep-Learning-Frameworks auf der Intel® Architektur**. Software-Entwickler, die sich für das Thema Deep Learning interessieren, haben möglicherweise schon von Intel MKL-DNN gehört, aber noch nicht die Gelegenheit gehabt, es selbst auszuprobieren.
Intel® Distribution for Python*	Dieses Tool verbessert die Leistung der beliebtesten und am schnellsten wachsenden Programmiersprache, die damit beinahe native Geschwindigkeit erreicht. Das Tool bietet einen schnellen und einfachen Zugang zu High-Performance-Python* mit Drop-In-Ersatz, vielen Optimierungstechniken und schnellem Zugriff auf Intel® Architekturoptimierungen.

Vorhersagemodell zur Bestimmung des Knochenalters

188-FACHE STEIGERUNG

in Bildern pro Sekunde[4]

Lungensegmentierungsmodell

38-FACHE STEIGERUNG

in Bildern pro Sekunde4

PRAXISBEISPIEL:

PHILIPS*: EFFIZIENTE, KI-GESTÜTZTE MEDIZINISCHE BILDGEBUNG

Intel und Philips wollten gemeinsam zeigen, dass Server mit skalierbaren Intel® Xeon® Prozessoren dazu genutzt werden können, Deep-Learning-Inferenzen von Röntgenbildern und CT-Aufnahmen von Patienten ohne Beschleuniger effizient durchzuführen. Das vorrangige Ziel von Philips ist es, seinen Endkunden KI anzubieten, ohne die Kosten für deren Systeme deutlich zu erhöhen oder Modifizierungen an der in dem Bereich eingesetzten Hardware notwendig zu machen. Die Unternehmen haben zwei medizinische Anwendungsfälle von Deep-Learning-Inferenz-Modellen getestet: Röntgenbilder von Knochen zur Vorhersagemodellierung des Knochenalters sowie CT-Aufnahmen von Lungen zur Lungensegmentierung.

Erfahren Sie, welche Ergebnisse Philips erzielt hat

PRAXISBEISPIEL:

ZIVA DYNAMICS*: GRENZEN DER SIMULATION DURCH KI VERSCHIEBEN

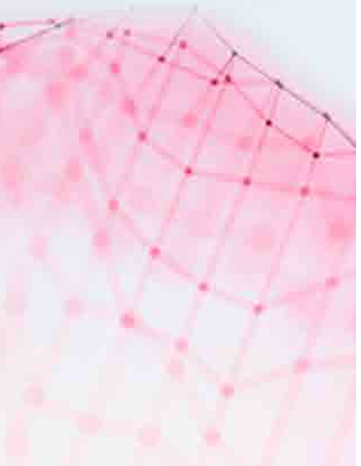
Mittels CGI erzeugte visuelle Effekte (VFX) verlangen im Allgemeinen ein komplexes Zusammenspiel von Know-how, Technologie und oft zeitaufwändigen kreativen Iterationen. Ziva Dynamics löst das durch Nutzung von Simulationssoftware, die auf künstlicher Intelligenz basiert. Dadurch können VFX-Künstler Figuren schaffen, die realistisch aussehen und sich den Gesetzen der Physik entsprechend korrekt bewegen. Ziva lässt seine Software in einer Umgebung laufen, die auf skalierbaren Intel® Xeon® Prozessoren aufgebaut ist. Der Großteil der Software wurde mit der Intel® Math Kernel Library (Intel® MKL) PARDISO* und dem Intel MKL Linear Algebra Package* (LAPACK*) geschrieben. Dadurch kann das Unternehmen schnell realistische Effekte erzeugen.

Lesen Sie den gesamten Bericht über Ziva Dynamics

6. NÄCHSTE SCHRITTE: ÜBERWINDEN DER KLUFT ZWISCHEN MODELL UND WIRKLICHKEIT

Die Auswirkungen von KI auf Technologie und Gesellschaft sind zwar noch recht neu, aber die Dynamik ist bereits spürbar. Einige führende Unternehmen und Märkte nutzen KI bereits. Jene jedoch, die erst jetzt damit beginnen, sollten darüber nachdenken, wie sie Deep-Learning-Inferenz schrittweise einführen können. Das beginnt mit der KI-Bereitschaft: Geschäftsszenarien untersuchen, Daten in Ordnung bringen und die richtige Mischung aus Personen und Technologie finden, die den KI-Hype in ihrem Unternehmen in die Realität umsetzen.

Bevor Sie über die ersten oder nächsten Schritte auf dem Weg zur KI nachdenken, sollte Sie ermitteln, an welcher Stelle im KI-Bereitschaftsmodell sich Ihr Unternehmen befindet. Je nach Bereitschaftsgrad sollte mit dieser Checkliste sichergestellt werden, dass Ihr Unternehmen mit den Geschäftszielen, Tools, Mitarbeitern und Sicherheitsüberlegungen wachsen kann, um erfolgreich zu sein.



Grundsätzliche Bereitschaft: Erste Nutzung von KI

- Ist das Szenario, der Anwendungsfall oder das mit KI zu lösende Problem klar definiert?
- Sind die Prioritäten so gesetzt, dass KI den größten geschäftlichen Nutzen bringt?
- Ist die geplante Infrastrukturarchitektur klar und zweckmäßig?
- Sind alle notwendigen Datenquellen klar ersichtlich und zugänglich?
- Können die von Ihnen gewählten Softwarepakete die KI-Lösung durchgängig umsetzen?
- Sind ausreichende Fähigkeiten und Ressourcen vorhanden (intern oder extern)?
- Wurden Erwartungen an die Schulungs- und Lernzeiten gestellt?
- Sind die Gesamtbetriebskosten (Total Cost of Ownership, TCO) der durchgängigen Lösung klar und genehmigt?

Einsatzbereitschaft: Skalierung der KI-Nutzung

- Kann die geplante Lösung über die ersten Tests und Evaluierungen hinaus skaliert werden?
- Wurde ein klar definiertes Geschäftsszenario mit dem Geschäftsbereich vereinbart?
- Sind genügend direkte Ressourcen verfügbar, mit zugeteilter und reservierter Zeit?
- Ist die Netzwerkbandbreite ausreichend, um eine zeitnahe Datenbereitstellung zu gewährleisten?
- Gibt es operative Managementprozesse, die die KI-Ergebnisse einbeziehen?
- Entspricht die Architektur den Industriestandards und bewährten Verfahren?
- Wurde eine Risikobewertung der Cybersicherheit vorgenommen und umgesetzt?
- Wurden realistische Umsetzungspläne formuliert und vermittelt?

Transformationsbereitschaft: Ausweitung der KI-Nutzung

- Gibt es ein Team, das die kontinuierliche Verbesserung der KI überwacht?
- Wurden die erweiterten KI-Möglichkeiten für das Unternehmen untersucht und sind sie bekannt?
- Wurden KI-Lösungen nach bewährten Verfahren entwickelt und eingesetzt?
- Gibt es Maßnahmen zur Überwachung der Geschäftseffektivität der KI-Lösungen?
- Wird die Architektur der KI als Plattform und nicht als Einzellösung realisiert?
- Sind die jeweiligen Unternehmensbereiche umfassend über die Auswirkungen der KI auf ihre Prozesse informiert?
- Werden die Governance-Bedürfnisse der KI-Lösung klar verstanden?
- Wird KI als zentrale Säule einer IT-gestützten Geschäftsstrategie gesehen?

WEITERE INFORMATIONEN

- **Erfahren Sie mehr: Intel® Deep Learning Boost >**
- **Whitepaper: Deep-Learning-Inferenz und -Training mit geringerer numerischer Genauigkeit >**
- **Solution Brief: Intel Select Solutions für Big DL Apache Spark >**
- **Solution Brief: Intel Select Solutions für das KI-Inferencing >**

Durch Technologien von Intel ermöglichte Funktionsmerkmale und Vorteile hängen von der Systemkonfiguration ab und können entsprechend geeignete Hardware, Software oder die Aktivierung von Diensten erfordern. Die Leistung kann je nach Systemkonfiguration unterschiedlich ausfallen. Kein Computersystem bietet absolute Sicherheit. Informieren Sie sich beim Systemhersteller oder Fachhändler oder auf intel.de.

In Leistungstests verwendete Software und Workloads können speziell für die Leistungseigenschaften von Intel® Mikroprozessoren optimiert worden sein. Leistungstests wie SYSmark* und MobileMark* werden mit spezifischen Computersystemen, Komponenten, Softwareprogrammen, Operationen und Funktionen durchgeführt. Jede Veränderung bei einem dieser Faktoren kann abweichende Ergebnisse zur Folge haben. Als Unterstützung für eine umfassende Bewertung Ihrer geplanten Anschaffung sollten Sie noch andere Informationen und Leistungstests heranziehen – auch im Hinblick auf die Leistung des betreffenden Produkts in Verbindung mit anderen Produkten. Ausführlichere Informationen finden Sie unter http://www.intel.de/benchmarks.

Die Leistungsergebnisse basieren auf Tests, die zu dem in den Konfigurationen angegebenen Datum durchgeführt wurden, und spiegeln möglicherweise nicht alle öffentlich erhältlichen Sicherheitsupdates wider. Weitere Einzelheiten finden Sie in den veröffentlichten Konfigurationsdaten. Kein Produkt und keine Komponente bietet absolute Sicherheit.

[1] https://www.nextplatform.com/2018/10/18/deep-learning-is-coming-of-age/

[2] 14-fache Inferenzdurchsatz-Verbesserung mit Intel® Xeon® Platinum 8280 Prozessor mit Intel® Deep Learning Boost: Von Intel getestet am 20.2.2019. Dual-Sockel Intel® Xeon® Platinum 8280 Prozessor, 28 Kerne, HT aktiviert, Turbo aktiviert, insgesamt 384 GB Arbeitsspeicher (12 Steckplätze, je 32 GB, 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x200004d), Ubuntu 18.04.1 LTS, Kernel 4.15.0-45-generic, SSD: 1 x INTEL SSDSC2BA80 (sda) – SSD mit 745,2 GB, INTEL SSDPE2KX040T7 (nvme1n1) – SSD mit 3,7 TB, Deep Learning Framework: Intel® Optimierung für Caffe*, Version: 1.1.3 (Commit-Hash: 7010334f159da247db3fe3 a9d96a3116ca06b09a), ICC-Version 18.0.1, MKL-DNN-Version: 0.17 (Commit-Hash: 830a10059a018cd2634d94195140cf2d8790a75a, Modell https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS = 64, syntheticData, 4 Instanzen/Dual-Sockel, Datentyp: INT8; Vergleich mit: Von Intel getestet am 11. Juli 2017: Dual-Sockel Intel® Xeon® Platinum 8180 Prozessor (2,50 GHz, 28 Kerne), HT deaktiviert, Turbo deaktiviert, Scaling-Governor festgelegt auf „Performance“ über intel_pstate-Treiber, 384 GB DDR4-2666-ECC-RAM. CentOS Linux, Release 7.3.1611 (Kern), Linux-Kernel 3.10.0-514.10.2.el7. x86_64. SSD: Intel® SSD der Produktreihe DC S3700 (800 GB, 2,5“, 6-GBit/s-SATA, 25-nm-Technik, MLC).

Leistung gemessen mit: Umgebungsvariable: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPUFreq festgelegt mit: cpupower frequency-set-d 2,5G -u 3,8G -g Performance. Caffe: (http://github.com/intel/caffe/), Revision f96b759f71b2281835f690 af267158b82b150b5c. Inferenz gemessen mit „caffe time-- forward_only“-Befehl, Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein synthetischer Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Systemspeicher zwischengespeichert. Topologie-Spezifikationen von https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ Compiler, Version 17.0.2 20170213, Intel MKL Small Libraries, Version 2018.0.20170425. Caffe ausgeführt mit „numactl -l“.

[3] 30-fache Inferenzdurchsatz-Verbesserung mit Intel® Xeon® Platinum 9282 Prozessor mit Intel® Deep Learning Boost: Von Intel getestet am 26.2.2019. Plattform: Dual-Sockel „Dragon Rock“ mit Intel® Xeon® Platinum 9282 Prozessor (56 Kerne pro Prozessor), HT aktiviert, Turbo aktiviert, insgesamt 768 GB Arbeitsspeicher (24 Steckplätze, je 32 GB, 2933 MHz), BIOS: SE5C620.86B.0D.01.0241.112020180249, CentOS* 7 Kernel 3.10.0-957.5.1.el7. x86_64, Deep Learning Framework: Intel® Optimierung für Caffe*, Version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL-DNN-Version: 0.17 (Commit-Hash: 830a10059a018cd-2634d94195140cf2d8790a75a, Modell https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS = 64, keine Datenebene – syntheticData: 3x224x224, 56 Instanzen / Dual-Sockel, Datentyp: INT8; Vergleich mit: Von Intel getestet am 11. Juli 2017: Dual-Sockel Intel® Xeon® Platinum 8180 Prozessor (2,50 GHz, 28 Kerne), HT deaktiviert, Turbo deaktiviert, Scaling-Governor festgelegt auf „Performance“ über intel_pstate-Treiber, 384 GB DDR4-2666-ECC-RAM. CentOS Linux, Release 7.3.1611 (Kern), Linux-Kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD der Produktreihe DC S3700 (800 GB, 2,5“, 6-GBit/s-SATA, 25-nm-Technik, MLC).

Leistung gemessen mit: Umgebungsvariable: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=5 6, CPUFreq festgelegt mit: cpupower frequency-set-d 2,5G -u 3,8G -g Performance. Caffe: (http://github.com/intel/caffe/), Revision f96b759f71b2281835f690 af267158b82b150b5c. Inferenz gemessen mit „caffe time-- forward_only“-Befehl, Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein synthetischer Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Systemspeicher zwischengespeichert. Topologie-Spezifikationen von https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ Compiler, Version 17.0.2 20170213, Intel MKL Small Libraries, Version 2018.0.20170425. Caffe ausgeführt mit „numactl -l“.

[4] Konfigurationsdetails: Hardware: Intel® Xeon® Platinum 8168 Prozessor mit 2,70 GHz, Intel® Hyper-Threading-Technik (Intel® HT-Technik) deaktiviert. BIOS Version: SE5C620.86B.OD.01.0010.072020182008. Systemspeicher: 192 GB, 2,666 MHz. Intel® Turbo-Boost-Technik aktiviert. SSD: ATA-Gerät mit nicht wechselbarem Datenträger, Modellnummer: INTEL SSDSCS2CW240A3. Software: Ubuntu 18.04.1 LTS (GNU/Linux 4.15.0-29-generic x86_64*. Keras 2.1.1. TensorFlow 1.2.1. OpenVINO™ Toolkit 2018 R2. Intel® Math Kernal Library für Deep Neural Networks v0.14. Datensätze: Vorhersagemodell zur Bestimmung des Knochenalters: 299x299x3 .png-Bilder. Lungensegmentierungsmodell: 512x512 .dcm-Bilder.

Hinweise zur Optimierung: Unter Umständen können Intel-Compiler bei Optimierungen, die nicht für Mikroprozessoren von Intel spezifisch sind, auch bei Mikroprozessoren anderer Hersteller denselben Optimierungsgrad erzielen. Zu diesen Optimierungen gehören Befehlssätze für SSE2, SSE3 und SSSE3 sowie weitere Optimierungen. Intel übernimmt keine Garantie für die Verfügbarkeit, Funktionalität oder Wirksamkeit von Optimierungen für Mikroprozessoren, die nicht von Intel hergestellt wurden. Mikroprozessor-abhängige Optimierungen in diesem Produkt sind für Anwendung in Verbindung mit Intel-Mikroprozessoren bestimmt. Bestimmte, nicht für die Intel-Mikroarchitektur spezifische Optimierungen sind für Intel-Mikroprozessoren reserviert. Entnehmen Sie weitere Informationen zu den spezifischen Befehlssatzerweiterungen, auf die dieser Hinweis zutrifft, bitte den entsprechenden Benutzer- und Referenzhandbüchern.

Revisionshinweis: #20110804

Die Ergebnisse wurden unter Verwendung interner Analysen oder Architektursimulationen bzw. -modellen von Intel geschätzt oder nachempfunden. Sie dienen nur zu Informationszwecken. Unterschiede in der Hardware, Software oder Konfiguration des Systems können die tatsächliche Leistung beeinflussen.

Intel hat keinen Einfluss auf und keine Aufsicht über die Daten Dritter. Sie sollten diese Inhalte prüfen, andere Quellen heranziehen und sich davon überzeugen, dass die angeführten Daten zutreffen.

Alle hierin gemachten Angaben können sich jederzeit ohne besondere Mitteilung ändern. Wenden Sie sich an Ihren Ansprechpartner bei Intel, um die neuesten Produktspezifikationen und Roadmaps zu erhalten.

0519/CAT/MMK/PDF

338678-002DE