Machine learning on resource-constrained microcontrollers for edge AI and IoT applications

Figure 1: At IMMS, algorithms for embedded systems were designed, optimised and implemented in a demonstrator for AI-based fan monitoring using the machine learning approach TinyML: Motor bearings or fan blades can be directly monitored with the compact retrofittable solution. Defective parts are indicated directly via a red LED. Photograph: IMMS.

Motivation and Overview

For many industrial applications, distributed sensor technology is required at various points and often over a long period of time directly in the process, e.g. to collect data for predictive maintenance or wear or to detect fault conditions as well as anomalies in the systems. IMMS is researching on adaptive edge AI systems to integrate AI directly into the sensor and thus make decisions in real time without a detour via the cloud. Automated adaption to new environmental conditions opens up versatile edge computing and IoT applications. To achieve this, IMMS is working on one hand towards making embedded ultra-low-power systems (ULP) more and more energy-efficient. On the other hand, IMMS is researching on developing and optimising machine-learning algorithms and software for such ULP systems.

The reason for this is that these miniature devices operate with a power consumption of only a few milliwatts and only a few kilobytes of internal memory and, due to their limited resources, cannot be used for conventional machine learning models designed for high-performance computers. To obtain the results needed for

Lead application
Adaptive edge AI systems for industrial application:
www.imms.de

Core topic
Embedded AI:
www.imms.de

the application with as few resources as possible, IMMS has built its solutions on Tiny ML (tiny machine learning). This branch of machine learning includes hardware, algorithms and software. With it, IMMS was able to run AI algorithms directly on the sensor and thus evaluate and analyse the data directly on battery-powered ULP systems.

The example shown here can be used to monitor the condition and maintenance requirements of fans. This can also save energy, as no data transmission to an energy-intensive server is required. In addition, the security of the data is guaranteed as it is processed directly at the sensor and no raw data is transmitted.

The aim is to use these solutions as a basis for developing adaptive edge AI systems for further industrial applications. As an example, classical ML algorithms and deep neural networks are being designed, developed and trained at IMMS, which, beyond pure data collection and analysis, also enable AI-based predictions for predictive maintenance applications, such as prognosis of remaining useful lifetimefor drill bits.

Lead application
Adaptive edge AI systems for industrial application:
www.imms.de

Classical Machine Learning versus Deep Learning

When designing AI-algorithms for predictive maintenance or condition analysis of machines, two basic approaches can be used. The classic machine learning with feature engineering and a deep learning approach.

Classical machine learning involves algorithms that recognise patterns from processed and structured data in a goal-oriented and automated manner. Without a defined objective and processed data, no analyses and predictions are possible. Machine learning is therefore generally used for small, structured data sets.

For the applications mentioned, relevant information, characteristics or features are calculated from the sensor signals by means of signal processing. For this purpose, statistical methods can be used for analysis in the time domain, such as mean value, kurtosis or crest factor, but also analyses in the frequency domain, such as Fourier or wavelet transforms.

Core topic
Embedded AI:
www.imms.de

In **deep learning**, on the other hand, a large amount of unstructured data is processed in many iterations, analogous to human learning. Artificial neural networks

extract the features and structures required for analysis from the data themselves. In contrast to machine learning, data processing and feature extraction are carried out autonomously, if enough data and computing power are available.

For the aforementioned applications with deep learning, the sensor data are fed directly into a neural network, an AI model. For the training of such a model, a lot of data is needed, distributed as equally as possible for different machine states.

Machine learning, on the other hand, is more complex than the deep learning approach, but can be partially explained to the developer by the manual selection of features. The machine learning approach is more suitable when expert knowledge can be used or when only few data is available, as is often the case in predictive maintenance.

Optimisation in classic machine learning

High-dimensional AI models are powerful but require memory, computing power and energy

AI models are often designed to achieve maximum classification or prediction accuracy. Optimisation is one of the most important aspects of AI-based intelligent algorithms in order to implement them on resource-constrained, energy-efficient embedded systems. System requirements such as energy consumption and hardware costs depend on the model used. For predictive maintenance (PdM) and machine health estimation, not only raw sensor data but also spectral information statistics are very informative, creating a very large design space. Each additional sensor and feature extractor improves the performance of the AI model, but also increases the system complexity. Such a large feature space increases the system requirements in terms of memory, computing power and energy. This makes PdM on resource-constrained devices such as microcontrollers a major challenge.

Feature Ranking streamlines AI models by sorting out redundant features

The following investigations were carried out on the publicly available datasets Prognostia[1] and XJTU[2] in order to achieve comparability of results for existing models. These data sets contain sensor data on the wear process of ball bearings.

Lead application
Adaptive edge AI systems for industrial application:
www.imms.de

Core topic
Embedded AI:
www.imms.de

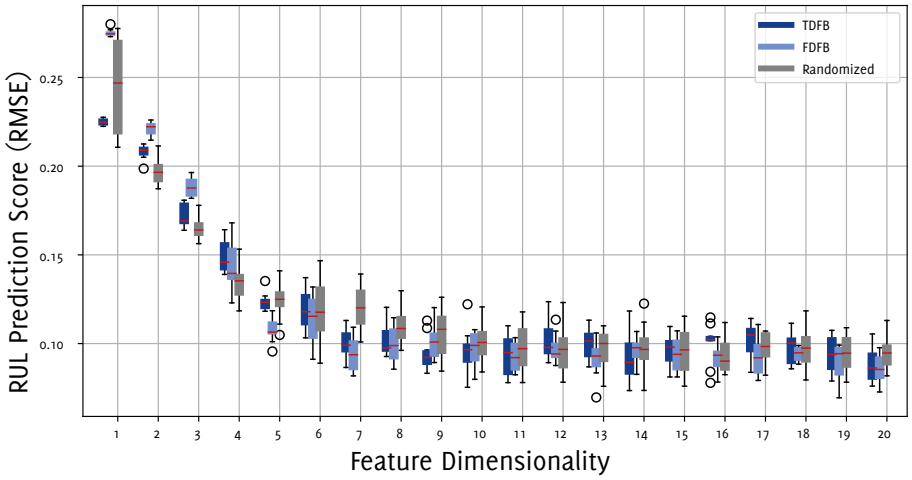

Figure 2: Attributes sorted by means of feature ranking. Graphic: IMMS.

In many applications, each additional sensor information or feature extractor can potentially improve the overall performance of the AI model. However, more features and thus higher dimensionality do not always guarantee higher prediction accuracy. Highly correlated features and overfitting can lead to a suboptimal solution.

Lead application
Adaptive edge AI systems for industrial application:
www.imms.de

One way to solve the high dimensionality problem is to use feature ranking (FR) algorithms to analyse the effects of each variable (feature or input) that are fed into the AI model. FR methods provide a detailed, interpretable metric of feature importance which can be used to build an optimised system for the particular application. There are a few different approaches that have been studied and evaluated at IMMS: wrapper methods, which are computationally intensive but very accurate; filter methods, which are very fast to compute, but can deviate from the optimal results; embedded methods, which require one-time training and are more accurate compared to filter methods.

Core topic
Embedded AI:
www.imms.de

The aim of the work was to find the optimal FR method for analysing the information content of features or input vectors, in order to reduce the overall system resources (computational effort, memory, energy requirements). To do this, the non-informative, redundant input signals and their features are removed from the AI model. An example is shown in figure 2. The x-axis represents the dimension of the features coming from two different vibration sensors (horizontal and vertical axes), which have been split into their spectral components. The y-axis is the error of the trained AI model in state estimation. Each box plot illustrates the error statistics for

the given dimension of features. It can be observed that the estimation error does not improve significantly further after 7 features. This means that the remaining 13 out of 20 features are not meaningful, which may be due to noise and redundant information.

For the example considered, a publicly available dataset, most of the informative features come from one sensor. This means that hardware and computational costs can be reduced by completely removing the other sensor and its corresponding features from the system. This reduced feature dimensionality can serve as a new basis for the development of optimised embedded systems.

Optimisation for Deep Learning for resource-constrained systems

Services for the development of embedded systems: www.imms.de

Neural networks can be reduced or made smaller by pruning. This involves removing connections and entire neurons in the network. The aim of pruning is to reduce the size of the network in such a way that the accuracy of the pruned model is only minimally reduced.

The approaches can be roughly divided into structured and unstructured pruning. Both have their advantages and disadvantages. Neural networks can also be represented as matrices. Unstructured pruning has only a minor effect on the accuracy of the pruned model, but can force operations on sparse matrices that are difficult to accelerate. This can lead to longer processing times and thus higher energy consumption of the system.

Core topic Embedded AI: www.imms.de

IMMS has been working on structured pruning by investigating and optimising AlexNet andResNet AI models based on a publicly available data set (XJTU[2]) and the results of other research projects.[3] Structured pruning does not lead to sparse matrices and is better suited for implementation on microcontrollers. In the models, structured pruning removed the complete filters of the convolutional layers of the model.. This can change the shape of the inputs and outputs of the layers but allows operations on dense matrices. Structured pruning can have a significant impact on the achieved accuracy if done aggressively. Therefore, a threshold value of 3% for the maximum difference in accuracy between the original AlexNet and ResNet models and the pruned models for the accuracy was set before pruning. Pruning is stopped when the accuracy difference falls below3%.

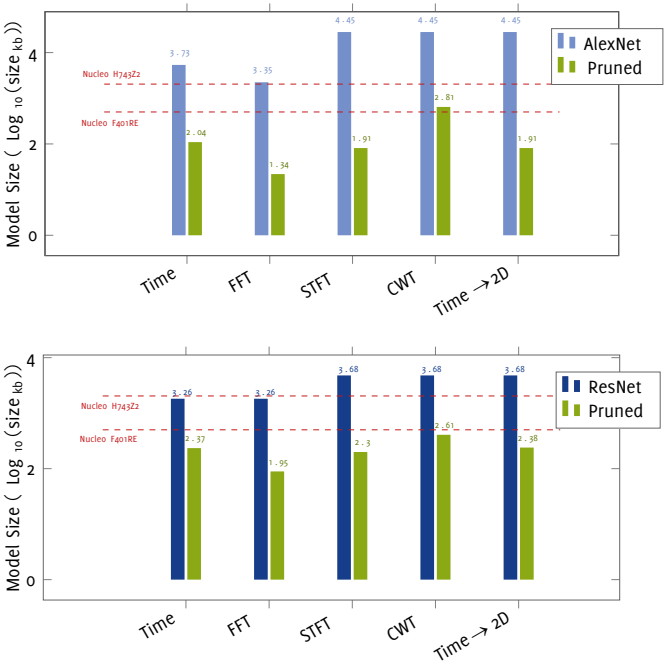

Figure 3:

Comparison of the memory requirements of the original AlexNet and ResNet models with the pruned models, and illustration of the available memory of two microcontroller types (STM32F401 and STm32H743).

Diagrams: IMMS.

Figure 3 shows the results of the memory requirements of the original models and the pruned models. For better classification, the available memory of two microcontroller types (STM32F401 and STM32H743) are also shown. The original models require more memory than is available in the microcontroller. In contrast, the adapted models can be implemented on one microcontroller. The table compares the achieved accuracies of the models.

Core topic Embedded AI: www.imms.de

	AlexNet		ResNet	
Features	**Original**	**Pruned**	**Original**	**Pruned**
Time signal	96,10%	93,23%	99,73%	97,02%
FFT Fast Fourier transform	98,4%	96,87%	99,21%	97,84%
CWT Continuous wavelet transform	90,02%	87,43%	92,91%	90,23%
2D time signal	97,70%	95,35%	95,57%	93,23%
STFT Short-time Fourier transform	98,37%	97,65%	99,6%	96,35%

Demonstrator

IMMS has created a demonstrator for AI-based fan monitoring in order to test the methods presented for the optimised use of algorithms for embedded ULP systems and to illustrate them in the application field of predictive maintenance, see figure 1. Motor bearings or fan blades can be monitored directly with the compact retrofit solution. The fans can be switched on and off individually on the demonstrator. The sensor system is placed in the centre of one of the fans and is attached by a magnet. Defective parts are indicated directly via an LED in red, intact ones in green.

First of all, a data set was recorded using ten commercially available PC fans with vibration sensors. The ball bearings of three fans were worn out. Using the two approaches presented, i.e. classical ML and a neural network, two AI models were trained with the fan dataset which can estimate the state (intact and defective) of the fans.

A battery-powered system consisting of a vibration sensor, microcontroller (STM32L4) and LEDs for status display were developed, manufactured and put into practice. The two AI models were optimised using the methods presented and implemented on the microcontroller.

Core topic Embedded AI: www.imms.de

Applications

In its work on adaptive edge AI systems, IMMS focuses on determining and forecasting machine and tool conditions for predictive maintenance applications with the help of AI. In addition to the example application for AI-based fan monitoring described above, IMMS has retrofitted vibration sensors to a metal-cutting machine in a further application and recorded the vibrations occurring during drilling at several points in the machining area. The aim of the investigation was to estimate the remaining service life of the drills in order to thus make optimum use of the tools and to avoid rejects due to damaged tools. To reduce the amount of data and derive specific features, the data is first pre-processed with various signal processing operations. With the data from several measurement series, the AI model was trained on the basis of an artificial neural network. The result can be displayed via a compact box retrofitted directly to the machine, both as a predicted remaining tool life as well as a classified wear condition of the tool in real time.

Lead application Adaptive edge AI systems for industrial application: www.imms.de

Summary and outlook

IMMS has successfully implemented the methods for optimising AI models for ULP systems in a demonstrator and published as well as presented them at technical conferences. In the future, the focus will be on the automatic analysis and selection of the important features, where currently more complex manual interventions are still necessary. In addition, IMMS will investigate how the pruning of neural networks can be applied to other AI models, e.g. autoencoders. The demonstrator will soon be extended with a radio interface for integration into existing systems. The aim of further developments is to make AI solutions more applicable for SMEs by automating the time-consuming processes involved in AI development as far as possible. This should significantly accelerate the derivation of further applications.

Contact person: Dipl.-Ing. Sebastian Uziel, sebastian.uziel@imms.de

Literature:

[1] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Chebel-Morello, N. Zerhouni, and C. Varnier, **Pronostia: An experimental platform for bearings accelerated degradation tests.** *in IEEE International Conference on Prognostics and Health Management, PHM'12. IEEE Catalog Number: CPF12PHM-CDR, 2012, pp. 1 – 8.*

[2] Biao Wang, Yaguo Lei, Naipeng Li, Ningbo Li, **A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings.** *IEEE Transactions on Reliability 2020, 69, 401 – 412.*

[3] Zhibin Zhao, Tianfu Li, Jingyao Wu, Chuang Sun, Shibin Wang, Ruqiang Yan, Xuefeng Chen: **Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study.** *ISA Transactions 2020, 107, 224 – 255.*

The research on methods for the optimised use of algorithms in adaptive edge AI systems was developed in the internal AI research group which is funded by the Land of Thüringen. The work on the example applications or the demonstrator was funded in the "SME Digital Centre Ilmenau" by the German Federal Ministry of Economics and Climate Protection (BMWK) under the reference 01MF21008C.

Supported by:

Federal Ministry for Economic Affairs and Climate Action

on the basis of a decision by the German Bundestag

Mittelstand-Digital

Mittelstand-Digital Zentrum Ilmenau

Services for the development of embedded systems: www.imms.de

Lead application Adaptive edge AI systems for industrial application: www.imms.de

Core topic Embedded AI: www.imms.de

Maschinelles Lernen auf ressourcenbeschränkten Mikrocontrollern für Edge-KI- und IoT-Anwendungen

Abbildung 1: Am IMMS wurden mithilfe des Machine-Learning-Ansatzes TinyML Algorithmen für eingebettete Systeme entworfen, optimiert und in einem Demonstrator zur KI-basierten Lüfterüberwachung realisiert: Motorlager oder Lüfterschaufeln lassen sich mit der kompakten nachrüstbaren Lösung direkt überwachen. Defekte Teile werden über eine rote LED direkt angezeigt. Foto: IMMS.

Motivation und Überblick

Für viele Industrie-Anwendungen wird verteilte Sensorik benötigt, die an verschiedenen Stellen oft über lange Zeit direkt am Prozess z.B. Fehlerzustände oder Verschleiß detektiert oder Daten zur vorausschauenden Instandhaltung erhebt. Das IMMS forscht an adaptiven Edge-KI-Systemen, um Künstliche Intelligenz direkt in den Sensor zu integrieren und somit Entscheidungen in Echtzeit ohne Umweg über die Cloud abzuleiten. Durch die automatisierte Anpassung an neue Umgebungsbedingungen lassen sich vielseitige Edge-Computing- und IoT-Anwendungen erschließen.

Hierfür arbeitet das IMMS zum einen daran, eingebettete Ultra-Low-Power-Systeme (ULP) immer energieeffizienter zu gestalten. Zum anderen forscht das IMMS daran, Machine-Learning-Algorithmen und Software für solche ULP-Systeme zu entwerfen und zu optimieren. Denn diese Kleinstgeräte arbeiten mit einer Leistungsaufnahme von nur wenigen Milliwatt und mit nur wenigen Kilobyte Arbeitsspeicher und können aufgrund ihrer begrenzten Ressourcen für die herkömmlichen, für hochperformante Rechner ausgelegten Machine-Learning-Modelle nicht genutzt werden. Um mit möglichst wenig Ressourcen die für die Anwendung benötigten Ergebnisse zu erhalten,

Leitanwendung
Adaptive Edge-KI-Systeme für industrielle Anwendungen:
www.imms.de

Kernthema
Eingebettete KI:
www.imms.de

hat das IMMS seine Lösungen auf Tiny ML (tiny machine learning) aufgebaut. Dieses Teilgebiet des maschinellen Lernens beinhaltet Hardware, Algorithmen und Software. Damit konnte das IMMS KI-Algorithmen direkt am Sensor ausführen und somit die Daten direkt auf batteriebetriebenen ULP-Systemen auswerten und analysieren.

Im gezeigten Beispiel lässt sich damit der Zustand und der Wartungsbedarf von Lüftern ermitteln. Damit kann zudem Energie eingespart werden, da keine Datenübertragung zu einem energieintensiven Server notwendig ist. Darüber hinaus ist die Sicherheit der Daten gewährleistet, da sie direkt am Sensor verarbeitet und keine Rohdaten übertragen werden.

Ziel ist es, auf der Grundlage dieser Lösungen Entwicklungen für adaptive Edge-KI-Systeme für weitere industrielle Anwendungen abzuleiten. Beispielsweise werden am IMMS ML-Algorithmen entworfen und neuronale Netzwerke trainiert, die über die reine Datenerhebung und -auswertung hinaus auch KI-basierte Vorhersagen zur vorausschauenden Wartung ermöglichen, wie z.B. zur Vorhersage von Bohrerstandzeiten.

Leitanwendung Adaptive Edge-KI-Systeme für industrielle Anwendungen: www.imms.de

Klassisches Maschine Learning versus Deep Learning

Beim Entwurf von KI-Algorithmen für die vorausschauende Wartung bzw. zur Zustandsanalyse von Maschinen können grundsätzlich zwei Ansätze verwendet werden. Das klassische Machine-Learning mit Merkmalsextraktion, dem Feature Engineering, und ein Deep-Learning-Ansatz.

Beim **klassischen Machine-Learning** geht es um Algorithmen, die zielgerichtet und automatisiert aus vorbereiteten und strukturierten Daten Muster erkennen. Ohne definiertes Ziel und aufbereitete Daten sind somit keine Analysen und Vorhersagen möglich. Daher wird Machine Learning in der Regel für kleine, strukturierte Datenmengen eingesetzt.

Für die genannten Anwendungen werden mittels Signalverarbeitung relevante Informationen, Merkmale oder Features aus den Sensorsignalen berechnet. Dafür können statistische Methoden für die Analyse im Zeitbereich, wie z.B. Mittelwert, Kurtosis oder Crestfaktor, aber auch Analysen im Frequenzbereich, z.B. Fourier- oder Wavelet-Transformationen verwendet werden.

Kernthema Eingebettete KI: www.imms.de

Im **Deep Learning** werden dagegen sehr viele sowie unstrukturierte Daten in vielen Iterationen analog zum menschlichen Lernen verarbeitet. Künstliche neurona-

le Netzwerke erzeugen die für Analysen benötigten Merkmale und Strukturen aus diesen Daten selbst. Im Gegensatz zum Machine Learning erfolgt eine selbständige Datenaufbereitung und Feature-Extraktion, sofern genug Daten und Rechenleistung vorhanden sind.

Für die genannten Anwendungen mit Deep-Learning werden die Sensordaten direkt in ein neuronales Netzwerk, ein KI-Modell, gegeben. Für das Training eines solchen Modells werden viele Daten möglichst gleich verteilt auf die unterschiedlichen Maschinenzustände benötigt.

Machine Learning ist dagegen aufwändiger als der Deep-Learning Ansatz, ist aber für den Entwickler durch die händische Auswahl der Merkmale teilweise erklärbar. Der Machine-Learning-Ansatz ist besser geeignet, wenn auf Expertenwissen zurückgegriffen werden kann oder aber nur wenige Daten verfügbar sind, wie das oft in der vorausschauenden Wartung der Fall ist.

Leitanwendung Adaptive Edge-KI-Systeme für industrielle Anwendungen: www.imms.de

Optimierung im klassischen Machine Learning

Hochdimensionale KI-Modelle sind leistungsfähig, benötigen aber Speicher, Rechenleistung und Energie

KI-Modelle werden oft so entworfen, dass die maximale Klassifikations- oder Vorhersagegenauigkeit erreicht wird. Die Optimierung ist einer der wichtigsten Aspekte von KI-basierten intelligenten Algorithmen, um sie auf ressourcenbeschränkten, energieeffizienten eingebetteten Systemen implementieren zu können. Die Systemanforderungen wie Energieverbrauch und Hardwarekosten hängen vom verwendeten Modell ab. Für die vorausschauende Wartung (Predictive Maintenance, PdM) und zur Schätzung des Maschinenzustands sind nicht nur die Roh-Sensordaten, sondern auch die Statistiken der spektralen Informationen sehr informativ, wodurch ein sehr großer Designraum entsteht. Jeder zusätzliche Sensor und Merkmalsextraktor verbessert die Leistung des KI-Modells, erhöht aber auch die Systemkomplexität. Ein solch großer Merkmalsraum erhöht die Systemanforderungen hinsichtlich Speicher, Rechenleistung und Energie. Das macht PdM auf ressourcenbeschränkten Geräten wie Mikrocontrollern zu einer großen Herausforderung.

Kernthema Eingebettete KI: www.imms.de

Feature Ranking verschlankt KI-Modelle durch Aussortieren redundanter Merkmale

Die nachfolgenden Untersuchungen wurden an den öffentlich verfügbaren Datensätzen Prognostia[1] und XJTU[2] durchgeführt, um eine Vergleichbarkeit der Ergebnisse

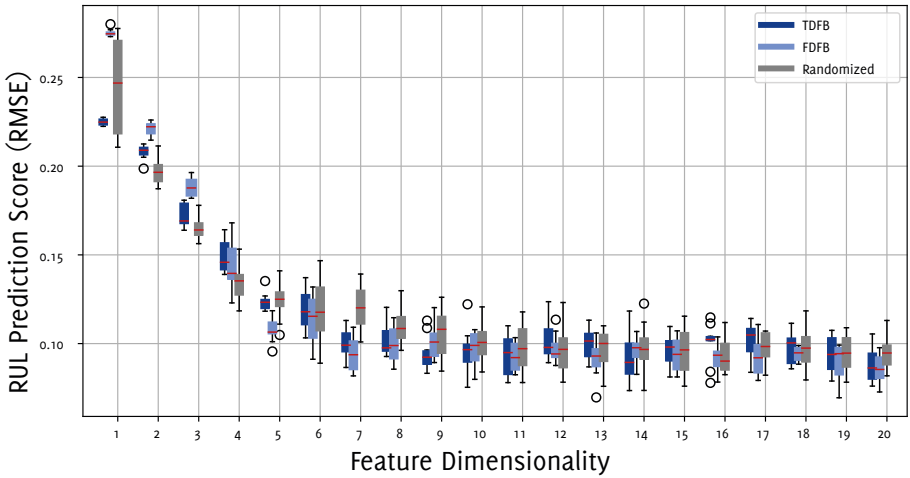

Abbildung 2: Mittels Feature Ranking geordnete Merkmale. Grafik: IMMS.

für vorhandene Modelle zu erzielen. Diese Datensätze enthalten Sensordaten zum Verschleißprozess von Kugellagern.

Bei vielen Anwendungen kann jede zusätzliche Sensorinformation oder jeder zusätzliche Merkmalsextraktor die Gesamtleistung des KI-Modells potenziell verbessern. Mehr Merkmale und damit eine höhere Dimensionalität garantieren jedoch nicht immer eine höhere Vorhersagegenauigkeit. Durch stark korrelierende Merkmale und Überanpassung kann es zu einer suboptimalen Lösung führen.

Eine Möglichkeit, das Problem der hohen Dimensionalität zu lösen, sind Algorithmen für das Feature-Ranking (FR), um die Auswirkungen jeder Variable (Feature oder Inputs) zu analysieren, die in das KI-Modell gespeist werden. FR-Methoden liefern eine detaillierte, interpretierbare Metrik der Merkmalsbedeutung, die für den Aufbau eines optimierten Systems für die jeweilige Anwendung verwendet werden kann. Es gibt einige verschiedene Ansätze, die am IMMS untersucht und evaluiert wurden: Wrapper-Methoden, die rechenintensiv, aber sehr genau sind; Filter-Methoden, die sehr schnell zu berechnen sind, jedoch von den optimalen Ergebnissen abweichen können; Embedded-Methoden, die ein einmaliges Training erfordern und im Vergleich zu Filter-Methoden genauer sind.

Ziel der Arbeiten war es, die optimale FR-Methode zur Analyse des Informationsgehalts von Merkmalen bzw. Eingangsvektoren zu finden, um somit die Gesamtsystemressourcen (Rechenaufwand, Speicher, Energiebedarf) zu senken. Dazu werden die nicht-informativen, redundanten Eingangssignale und ihre Merkmale aus dem KI-Modell entfernt. Ein Beispiel zeigt Abbildung 2. Die x-Achse stellt die Dimension

Leitanwendung Adaptive Edge-KI-Systeme für industrielle Anwendungen: www.imms.de

Kernthema Eingebettete KI: www.imms.de

der Merkmale dar, die von zwei verschiedenen Schwingungssensoren (horizontale und vertikale Achse) stammt und die in ihre spektralen Komponenten aufgeteilt wurden. Die Y-Achse ist der Fehler des trainierten KI-Modells bei der Zustandsschätzung. Jeder Box-Plot veranschaulicht die Fehlerstatistik für die gegebene Dimension an Features. Es ist zu beobachten, dass sich der Schätzfehler nach 7 Features nicht weiter signifikant verbessert. Dies bedeutet, dass die restlichen 13 von 20 Merkmalen nicht aussagekräftig sind, was auf Rauschen und redundante Informationen zurückzuführen sein kann.

Für das betrachtete Beispiel, einem öffentlich verfügbaren Datensatz, stammen die meisten informativen Merkmale von einem Sensor. Das bedeutet, dass die Hardware- und Rechenkosten reduziert werden können, indem der andere Sensor und die entsprechenden Merkmale aus dem System komplett entfernt werden. Diese reduzierte Merkmalsdimensionalität kann als neue Grundlage für die Entwicklung optimierter eingebetteter Systeme dienen.

Optimierung für Deep Learning für ressourcenbeschränkten Systeme

Neuronale Netze können durch Ausdünnen (engl. Pruning) reduziert bzw. verkleinert werden. Dabei werden Verbindungen und ganze Neuronen in dem Netzwerk entfernt. Ziel des Pruning ist es, das Netz so zu verkleinern, dass die Aussagegenauigkeit nur minimal sinkt.

Die Ansätze lassen sich grob in strukturiertes und unstrukturiertes Pruning unterteilen. Beide haben ihre Vor- und Nachteile. Neuronale Netze lassen sich auch als Matrizen darstellen. Unstrukturiertes Pruning wirkt sich nur geringfügig auf die Genauigkeit des beschnittenen Modells aus, kann aber Operationen auf schwachbesetzte Matrizen forcieren, die schwer zu beschleunigen sind. Dies kann zu längeren Abarbeitungszeiten und damit zu einem höheren Energieverbrauch des Systems führen.

Das IMMS hat sich bei seinen Arbeiten mit strukturiertem Pruning beschäftigt und dazu AlexNet- und ResNet-KI-Modelle basierend auf einem öffentlich verfügbaren Datensatz (XJTU[2]) und Ergebnissen anderer Forschungsarbeiten[3] untersucht und optimiert. Das strukturierte Pruning führt nicht zu dünnbesetzten Matrizen und eignet sich besser für die Implementierung auf Mikrocontrollern. Bei den Modellen wurden mit strukturiertem Pruning die kompletten Filter entfernt. Dies kann die Form der Ein- und Ausgänge der Schichten verändern, ermöglicht aber Operationen auf vollbesetzten Matrizen. Das strukturierte Pruning kann bei aggressivem Vorge-

Dienstleistung Entwicklung eingebetteter Systeme: www.imms.de

Kernthema Eingebettete KI: www.imms.de

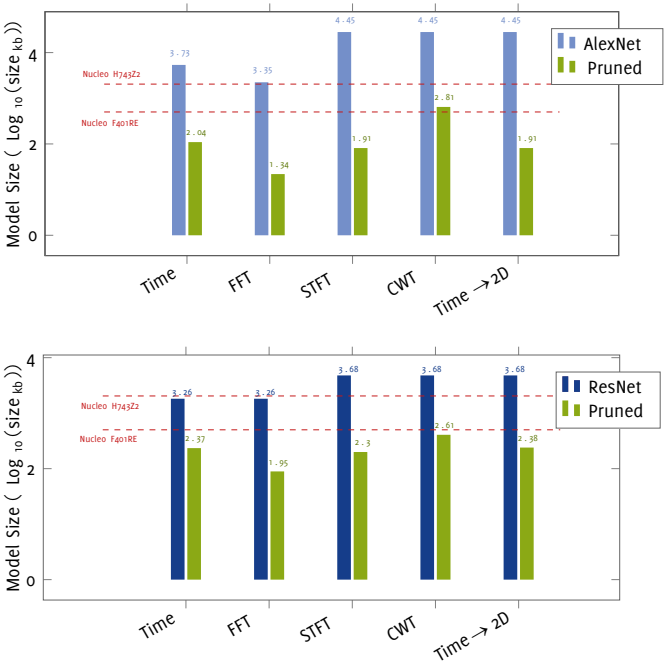

Abbildung 3:

Vergleich des Speicherbedarfs der ursprünglichen AlexNet- und ResNet- Modelle mit den beschnittenen Modellen, sowie Darstellung des verfügbaren Speichers von zwei Mikrocontroller-Typen (STM32F401 und STM32H743).

Grafiken: IMMS.

hen erheblichen Einfluss auf die erzielte Genauigkeit haben, daher wurde vor dem Pruning ein Schwellwert für den maximalen Genauigkeitsunterschied zwischen dem ursprünglichen AlexNet- und ResNet-Modell und dem beschnittenen Modell von 3% festgelegt. Das Pruning wird gestoppt, wenn der Genauigkeitsunterschied unter 3% fällt.

In der Abbildung 3 sind die Ergebnisse des Speicherbedarfs der ursprünglichen Modelle und der beschnittenen Modelle dargestellt. Zur besseren Einordung ist der verfügbare Speicher von zwei Mikrocontroller-Typen (STM32F401 und STM32H743) mit eingezeichnet. Die ursprünglichen Modelle benötigen mehr Speicher als im Controller vorhanden. Dem gegenüber können die angepassten Modelle auf einem Mikrocontroller implementiert werden. In der Tabelle sind die erzielten Genauigkeiten der Modelle gegenübergestellt.

Kernthema
Eingebettete KI:
www.imms.de

	AlexNet		ResNet	
Features	**Original**	**Pruned**	**Original**	**Pruned**
Zeitsignal	96,10%	93,23%	99,73%	97,02%
FFT Fast Fourier transform	98,4%	96,87%	99,21%	97,84%
CWT Continuous wavelet transform	90,02%	87,43%	92,91%	90,23%
2D Zeitsignal	97,70%	95,35%	95,57%	93,23%
STFT Short-time Fourier transform	98,37%	97,65%	99,6%	96,35%

Demonstrator

Das IMMS hat einen Demonstrator zur KI-basierten Lüfterüberwachung realisiert, um die vorgestellten Methoden zur optimierten Nutzung von Algorithmen für eingebettete ULP-Systeme zu erproben und im Applikationsfeld der vorausschauenden Wartung zu veranschaulichen, siehe Abbildung 1. Motorlager oder Lüfterschaufeln lassen sich mit der kompakten nachrüstbaren Lösung direkt überwachen. Die Lüfter können am Demonstrator einzeln an- und abgeschaltet werden. Das Sensorsystem wird mittig auf einem der Lüfter platziert und ist mittels eines Magneten befestigt. Defekte Teile werden direkt über eine LED rot angezeigt, intakte grün.

Zunächst wurde hierfür mit 10 handelsüblichen PC-Lüftern mit Vibrationssensoren ein Datensatz aufgenommen und erstellt. Bei drei Lüftern wurden die Kugellager verschlissen. Mit den zwei vorgestellten Ansätzen, dem klassischen ML und mit einem neuronalen Netzwerk wurden mit dem Datensatz zwei KI-Modelle trainiert, die den Zustand (intakt und defekt) der Lüfter abschätzen können.

Ein batteriebetriebenes System aus einem Vibrationssensor, Mikrocontroller (STM32L4) und LEDs zur Statusanzeige wurden entwickelt, gefertigt und in Betrieb genommen. Die beiden KI-Modelle wurden mit den vorgestellten Methoden optimiert und auf dem Mikrocontroller implementiert.

Kernthema Eingebettete KI: www.imms.de

Anwendungsfälle

Das IMMS fokussiert sich bei seinen Arbeiten zu adaptiven Edge-KI-Systemen darauf, Maschinen- und Werkzeugzustände zur vorausschauenden Wartung mithilfe von KI zu bestimmen und vorherzusagen. Neben der oben beschriebenen Beispielanwendung zur KI-basierten Lüfterüberwachung hat das IMMS dazu in einem weiteren Anwendungsfall Schwingungssensoren an einer Zerspannungsmaschine nachgerüstet und die während der Bohrung auftretenden Vibrationen an mehreren Stellen im Bearbeitungsraum erfasst. Ziel der Untersuchung war es, die Reststandzeit der Bohrer abzuschätzen, um somit die Werkzeuge optimal zu nutzen und Ausschuss durch beschädigte Werkzeuge zu vermeiden. Um die Datenmenge zu reduzieren und spezifische Merkmale abzuleiten, werden die Daten zunächst mit verschiedenen Signalverarbeitungsoperationen vorverarbeitet. Mit den Daten mehrerer Messreihen wurde das KI-Modell auf Basis eines künstlichen neuronalen Netzes trainiert. Das Ergebnis kann über eine direkt an der Maschine nachgerüstete kompakte Box sowohl als prognostizierte Reststandzeit als auch als klassifizierter Verschleißzustand des Werkzeugs in Echtzeit bereitgestellt werden.

Leitanwendung Adaptive Edge-KI-Systeme für industrielle Anwendungen: www.imms.de

Zusammenfassung und Ausblick

Das IMMS hat die Methoden zur Optimierung von KI-Modellen für ULP-Systeme erfolgreich in einem Demonstrator implementiert und publiziert bzw. auf Fachkonferenzen präsentiert. Der Fokus wird künftig in der automatischen Analyse und Auswahl der wichtigen Merkmale liegen, wo derzeit noch aufwändigere manuelle Eingriffe notwendig sind. Darüber hinaus wird das IMMS untersuchen, wie sich das Pruning von neuronalen Netzen auf andere KI-Modelle, z.B. Autoencoder, anwenden lässt. Der Demonstrator wird demnächst mit einer Funkschnittstelle zur Einbindung in bestehende Systeme erweitert. Ziel weiterer Entwicklungen ist, KI-Lösungen besser für KMU zu erschließen, indem aufwändige Prozesse bei der KI-Anwendungsentwicklung weitestgehend automatisiert werden. Das soll die Ableitung weiterer Anwendungen deutlich beschleunigen.

Kontakt: Sebastian Uziel, sebastian.uziel@imms.de

Literatur:

[1] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Chebel-Morello, N. Zerhouni, and C. Varnier, **Pronostia: An experimental platform for bearings accelerated degradation tests.** *in IEEE International Conference on Prognostics and Health Management, PHM'12. IEEE Catalog Number: CPF12PHM-CDR, 2012, pp. 1 – 8.*

[2] Biao Wang, Yaguo Lei, Naipeng Li, Ningbo Li, **A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings.** *IEEE Transactions on Reliability 2020, 69, 401 – 412.*

[3] Zhibin Zhao, Tianfu Li, Jingyao Wu, Chuang Sun, Shibin Wang, Ruqiang Yan, Xuefeng Chen: **Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study.** *ISA Transactions 2020, 107, 224 – 255.*

Die Forschung zu Methoden für die optimierte Nutzung von Algorithmen in adaptiven Edge-KI-Systemen wurden in der internen, vom Freistaat Thüringen finanzierten KI-Forschungsgruppe erarbeitet. Die Arbeiten zu den Beispielapplikationen bzw. der Demonstrator wurden im „Mittelstand-Digital-Zentrum Ilmenau" durch das Bundesministerium für Wirtschaft und Klimaschutz (BMWK) unter dem Kennzeichen 01MF21008C gefördert.

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

Mittelstand-Digital

Mittelstand-Digital Zentrum Ilmenau

Dienstleistung Entwicklung eingebetteter Systeme: www.imms.de

Leitanwendung Adaptive Edge-KI-Systeme für industrielle Anwendungen: www.imms.de

Kernthema Eingebettete KI: www.imms.de