

Masterarbeit

Computer Science

Sensorbasierter Orientierungssinn mit künstlichen neuronalen Netzen und Entscheidungsbäumen

von

Tom Dymel

Juli 2021

Betreut von

Dr. Marcus Venzke

Institute of Telematics, Hamburg University of Technology

Erstprüfer

Prof. Dr. Volker Turau

Institute of Telematics

Hamburg University of Technology

Zweitprüfer

Dr. ToDo

Institute of Telematics

Hamburg University of Technology

Inhaltsverzeichnis

1	Einleitung	1
2	Entscheidungsbäume	3
2.1	Scikit-Learn	3
2.2	Einzelne Entscheidungsbäume	4
2.3	Ensemble-Methoden	5
2.4	Training mit Scikit-Learn	7
2.5	Ressourcenbedarf auf dem Mikrocontroller	7
2.5.1	Ausführungszeit und Energieverbrauch	7
2.5.2	Programmgröße	8
3	Künstliche Neuronale Netze	11
3.1	Keras	13
3.2	Training von KNN	13
3.3	Lernalgorithmen	14
3.4	Aktivierungsfunktionen	16
3.5	Ressourcenbedarf auf dem Mikrocontroller	18
4	Standortbestimmung	21
4.1	Lokalisierung	21
4.2	Anwendung von Lokalisation	21
4.3	Standortbestimmung mit maschinellen Lernen	21
4.4	Orientierungssinn	21
5	Trainings- und Validationsdaten	23
5.1	Simulierten Sensordaten	23
5.2	Künstlichen Sensordaten	23
5.2.1	Magnetfeld	23
5.2.2	Temperatur	23
5.2.3	Lautstärke	24
5.2.4	WLAN Zugangspunkte	24
5.3	Simulation von Interrupts	24
5.4	Standortenkodierung	24
5.5	Feature-Extrahierung	24
5.6	Aufteilung der Daten	25
6	ML-Modelle	27
6.1	Entscheidungswald	27
6.2	Feed Forward Neuronales Netzwerk	27
6.3	Feedback Kanten	27
6.4	Training der Modelle	28

INHALTSVERZEICHNIS

7	Evaluation	29
7.1	Klassifizierungsgenauigkeit	29
7.2	Fehlertoleranz	29
7.3	Ressourcennutzung	30
8	Diskussion	31
9	Schlussfolgerungen	33
A	Inhalt des USB-Sticks	35
	Literaturverzeichnis	37

Einleitung

Lokalisation ist der Prozess die Position von einem Objekt zu bestimmen. Die Positionsbestimmung ist integral für viele technische System, z. B. Tracking, Navigation oder Überwachung. Ein sehr bekanntes und vielseitig genutztes Positionsbestimmungssystem ist das Global Positioning System (GPS). GPS trianguliert die Position des anfragenden Geräts mit Hilfe von mehreren Satelliten (TODO Quelle). Im freien ist eine Genauigkeit von $X(\text{TODO})$ m möglich. Für die Positionsbestimmung innerhalb von Gebäuden ist die Genauigkeit aber nicht ausreichend, außerdem wird sie erschwert durch die dicke Wände und Interferenzen. Dadurch ist GPS oft nicht ausreichend für Trackingsysteme innerhalb von Gebäuden, z. B. Lagerhallen. Aus diesem Grund werden andere Systeme für diesen Zweck verwendet. Je nach Bedarf der Genauigkeit können Objekte mit Sendern, RFID Tags oder Barcodes markiert werden (TODO Quellen). Diese Ansätze bedürfen eine Infrastruktur, die in den Gebäuden installiert und gewartet werden muss.

In dieser Arbeit wird die diskrete Positionsbestimmung basierend auf Sensordaten untersucht. Dabei soll eine bestimmte Anzahl an Orten anhand von verschiedenen Sensorwerten unterschieden werden. Dies ist Vergleichbar mit dem Orientierungssinn von Tieren und Menschen. Zum Beispiel navigieren Honigbienen auf Basis von gelernten Orientierungspunkten, um Nahrungsquelle und Nest zu finden (TODO Quelle). Mit Hilfe maschinellen Lernens sollen künstliche neuronale Netze (KNN) und Entscheidungsbäume trainiert werden. Als Eingabedaten werden aus den gesammelten Sensordaten Features extrahiert, d. h. Attribute und Eigenschaften dieser Daten. Dieses System bedarf keine Infrastruktur muss aber für jedes Gebäude individuell trainiert werden.

TODO: Was wurde in dieser Arbeit gemacht.

Kapitel 2 führt Entscheidungsbäume und Ensemble-Methoden ein. Kapitel 3 führt künstliche

1 EINLEITUNG

neuronale Netze (KNN) ein mit dem Fokus auf Feed Forward neuronale Netze (FFNN). In Kapitel 4 wird auf den Stand der Forschung für Standortbestimmung mit Hilfe von maschinellen Lernen (ML) eingegangen. Die Generierung von Trainings- und Validationsdaten auf Basis von Simulationen wird in Kapitel 5 erläutert. Kapitel 6 stellt die trainierten ML Modelle mit Entscheidungsbäumen und FFNN vor. Darauf folgt die Evaluation der Klassifizierungsgenauigkeit, Fehlertoleranz und Ressourcennutzung in Kapitel 7. Kapitel 8 enthält einen kritischen Rückblick auf die Entscheidungen dieser Arbeit, bevor Kapitel 9 Schlussfolgerungen zieht.

Entscheidungsbäume

Ein Entscheidungsbaum ist ein Baum mit dem Entscheidungen getroffen werden [Qui90]. Das geschieht, indem der Baum von der Wurzel zu einem Blatt traversiert wird. Dabei bestimmt ein Test in jedem inneren Knoten, mit welchem Kindknoten fortgefahren wird. Jedes Blatt entspricht einer Entscheidung des Entscheidungsbaums. Es wird unterschieden zwischen Bäumen, die versuchen eine der vordefinierten Klassen zu klassifizieren (Klassifizierer), und solchen, die versuchen den nächsten Wert vorherzusagen (Regressoren).

Die Konstruktion eines optimalen binären Entscheidungsbaumes ist NP-Vollständig [LR76]. Den optimalen Klassifizierer zu finden ist folglich sehr aufwendig. Aus diesem Grund werden bei der Konstruktion Heuristiken verwendet, die nur lokal die beste Entscheidung treffen. Zudem werden einzelne Entscheidungsbäume in einem Ensemble zu einem Entscheidungswald zusammengefasst, um den Fehler eines einzelnen Baumes zu reduzieren [Ent20b].

2.1 Scikit-Learn

Diese Arbeit verwendet die Python ML-Bibliothek *Scikit-Learn*. Scikit-Learn bietet verschiedene ML Algorithmen mit einem High-Level Interface an [PVG⁺11]. Zur Konstruktion der Entscheidungsbäume wird der Algorithmus von CART verwendet [Ent20a]. Die Bibliothek kann Klassifizierer und Regressoren generieren.

Relevant ist jedoch lediglich der Klassifizierer, da in dieser Arbeit ein Klassifizierungsproblem gelöst werden soll. Dieser bietet zahlreiche Hyperparameter an, um die Konstruktion des Entscheidungsbaumes zu steuern. In dieser Arbeit wird der Hyperparameter `max_depth` verwendet. Dieser Hyperparameter beschränkt die maximale Baumhöhe und begrenzt somit den Programmspeicherverbrauch.

Scikit-Learn bietet Ensemble-Methoden an, um Entscheidungswälder zu trainieren. Ein wichtiger Parameter dieser Methoden ist `n_estimators`. Dieser steuert die Größe des Ensembles bzw. die Waldgröße. Somit hat auch dieser Parameter Einfluss auf den Programmspeicherverbrauch.

2.2 Einzelne Entscheidungsbäume

Der einzelne Entscheidungsbaum ist eine rekursive Datenstruktur um Entscheidungsregeln darzustellen [Qui90]. Jeder innere Knoten ist ein *Test*, welcher eine arbiträre Anzahl von sich gegenseitig ausschließenden Ergebnissen hat. Das Ergebnis eines Tests bestimmt mit welchem Kindknoten fortgefahren wird. Die Blätter des Baumes stellen die Entscheidungen dar bzw. die Klassen des Entscheidungsbaumklassifizierers. Abbildung 2.1 zeigt einen binären Entscheidungsbaum, in dem jeder Test zwei mögliche Ergebnisse hat. Das Trainieren von



■ **Abbildung 2.1:** Beispiel eines binären Entscheidungsbaums mit 3 möglichen Ergebnissen.

Entscheidungsbäumen ist eine Art von *Supervised Learning*, d. h. aus einer beschrifteten Trainingsmenge werden Regeln abgeleitet, um das korrekte Mapping von Input zu Output abzubilden [GL09]. Die Trainingsmenge besteht aus Feature-Mengen, die mit Klassen beschriftet sind [Ste09]. Die Generalisierungsfähigkeit ist abhängig von der Trainingsmenge. Zum einen sollte die Trainingsmenge möglichst repräsentativ sein für die Aufgabe, die gelernt werden soll. Zum anderen sollten die verwendeten Features eine Partitionierung aller Klassen ermöglichen [PGP98].

Entscheidungsbäume werden heuristisch konstruiert, da die Konstruktion eines optimalen Entscheidungsbaumes NP-Vollständig ist [LR76]. Zu diesen Algorithmen gehören beispielsweise ID3 [Qui86], C4.5 [Qui14] oder CART [BFSO84]. Die Aufgabe ist durch gezielte Trennungen eine Partitionierung der Trainingsmenge zu erzeugen, sodass möglichst nur Einträge mit der gleichen Beschriftung in einer Partitionierung enthalten sind. Die Algorithmen unterscheiden sich in ihrer Strategie [Qui86].

Scikit-Learn implementiert eine optimierte Version des *CART* (Classification And Regression Trees) Algorithmus [Ent20a]. CART partitioniert die Trainingsmenge indem lokal immer die beste Teilung ausgewählt wird, d. h. es wird für die momentane Teilmenge immer die beste Teilungsregel ausgewählt. Dieser Vorgang wird rekursiv mit jeder Teilmenge wiederholt, bis keine weitere Teilung mehr möglich ist oder alle Einträge einer Partitionierung die gleiche Beschriftung tragen [Ste09].

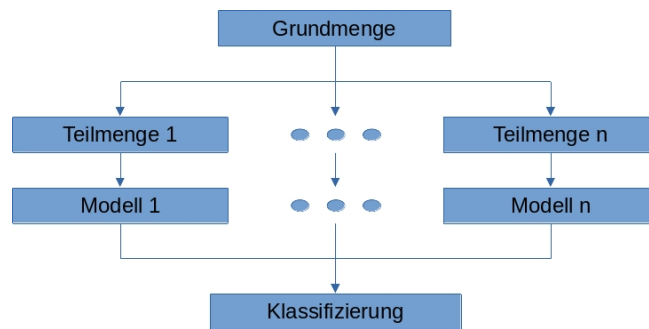
2.3 Ensemble-Methoden

Ensemble-Methoden beschreiben wie mehrere Entscheidungsbäume trainiert werden, um eine möglichst hohe Diversität der einzelnen Entscheidungsbäume zu erzielen. Das Ergebnis eines Ensembles ist die Aggregation der Ergebnisse der einzelnen Entscheidungsbäume [D⁺02].

Der Wahlklassifizierer $H(x) = w_1 h_1(x) + \dots + w_K h_K(x)$ ist eine Möglichkeit die Einzelergebnisse $\{h_1, \dots, h_K\}$ gewichtet mit $\{w_1, \dots, w_K\}$ zu aggregieren [D⁺02]. Ein Ergebnis kann auf zwei Arten modelliert sein. Einerseits als eine Funktion $h_i : D^n \mapsto \mathbb{R}^m$, die einer n -dimensionalen Menge D^n jeder der m möglichen Klassen eine Wahrscheinlichkeit zuweist. Das Ergebnis ist eine Wahrscheinlichkeitsverteilung. Das diskrete Ergebnis der Klassifizierung ist die Klasse mit der höchsten Wahrscheinlichkeit in dem Ergebnis. Andererseits kann es als eine Funktion $h_i : D^n \mapsto M$ abgebildet werden, die diskret auf eine der möglichen Klassen in M verweist [Dym21]. In diesem Fall wird die Klasse ausgewählt, die am häufigsten unter allen Einzelergebnissen vorkam. In der Praxis wird die Aggregation der Wahrscheinlichkeitsverteilung genutzt [Ent20b]. Analog ist $H : D^n \mapsto \mathbb{R}^m$ oder $H : D^n \mapsto M$ definiert [D⁺02]. Für gewöhnlich hat jeder Teilnehmer einer Wahl das gleiche Gewicht.

Bagging (**B**ootstrap **a**ggregating) konstruiert Entscheidungswälder, indem es Entscheidungsbäume mit Teilmengen der Trainingsmenge trainiert. Abbildung 2.2 illustriert die Bagging Methode für n Entscheidungsbaummodelle. Zunächst wird die Trainingsmenge in n Teilmengen aufgeteilt [Bre96]. Der Inhalt der Teilmengen wird mit der „Bootstrap sampling“ Methode bestimmt. Diese zieht aus einer Grundmenge l -mal jeweils k -Einträge [Efr92]. Mit jeder Teilmenge wird ein Entscheidungsbaum trainiert [Bre96]. Die Einzelergebnisse werden aggregiert, z. B. mit dem Wahlklassifizierer.

Random Forest erweitert die Bagging-Methode [Bre01]. Für jeden Entscheidungsbaum der trainiert werden soll, wird zusätzlich zufällig eine Teilmenge der Feature-Menge ausgewählt.



■ **Abbildung 2.2:** Klassifizierungsprozess mit der Bagging-Methode.

Extremely Randomized Trees (ExtraTrees) verwenden ebenfalls eine Teilmenge der Feature-Menge der Trainingsmenge beim Trainieren der einzelnen Entscheidungsbäume [GEW06]. Allerdings wird für jeden Entscheidungsbaum die gesamte Trainingsmenge verwendet. Dies soll den Bias reduzieren. Bei der Konstruktion wird nicht versucht die beste Teilungsregel zu finden, sondern es werden zufällig Teilungsregeln generiert, aus denen die Beste ausgewählt wird. Dieses Verfahren soll die Varianz reduzieren.

Beim Boosting werden nacheinander schwache Lerner auf einer Teilmenge trainiert, die gewichtet aggregiert werden [FS97]. Dadurch entsteht ein starker Lerner. Abbildung 2.3 illustriert, wie vier schwache Lerner trainiert werden. Jeder Lerner findet eine Funktion der die trainierte Teilmenge unterteilt. Anschließend werden sie gewichtet aggregiert. Dies konstruiert einen starken Lerner, der die gesamte Trainingsmenge unterteilt. Diese Arbeit verwendet AdaBoost [FS97] von Freund und Schapire.



■ **Abbildung 2.3:** Klassifizierungsprozess mit der Boosting-Methode.

2.4 Training mit Scikit-Learn

Die Konstruktion eines Entscheidungsbaumes mit Scikit-Learn ist nicht deterministisch [Dym21]. Bei der Konstruktion mit CART können zwei Teilungsregeln gefunden werden, die eine gleich gute Unterteilung erzeugen. In diesem Fall wählt Scikit-Learn zufällig eine Teilung aus. Dadurch werden anschließende Teilungen beeinflusst, die in einen der Fälle womöglich bessere Teilungen finden hätten können. Dieser Zufall ist steuerbar, indem der Startwert des Zufallgenerators auf einen vordefinierten Wert gesetzt wird.

Bei identischer Trainingsmenge und Konfiguration können folglich für verschiedene Startwerte unterschiedliche Modelle erzeugt werden. Aus diesem Grund kann das Training mit einem Startwert als Monte Carlo Methode verstanden werden, d. h. wiederholtes Ausführen unter verschiedenen Startwerten erhöht die Wahrscheinlichkeit, dass das beste Modell unter der angegebenen Konfiguration gefunden wird.

Dies bedarf, dass die Klassifizierungsgenauigkeit beim Training bereits ermittelt wird, damit das beste Modell ausgewählt werden kann. Dafür wird die Trainingsmenge in zwei Mengen unterteilt. Mit einer Teilmenge werden die Entscheidungsbäume unter verschiedenen Startwerten des Zufallgenerators trainiert und mit der anderen Teilmenge werden diese untereinander verglichen. Das Ergebnis des Trainingsprozesses ist das Modell, dass auf letzterer Teilmenge die höchste Klassifizierungsgenauigkeit erzielt.

2.5 Ressourcenbedarf auf dem Mikrocontroller

Zukünftig soll das Modell auf einem Mikrocontroller ausgeführt werden [Ven21]. Mikrocontroller sind stark limitiert in ihrer Rechenleistung, Speicherkapazität, RAM und werden oft zudem mit einer Batterie betrieben. Aus diesem Grund ist der Energieverbrauch zu minimieren und das Modell muss innerhalb dieser Limitierungen operieren können.

2.5.1 Ausführungszeit und Energieverbrauch

Der Energieverbrauch korreliert mit der Ausführungszeit. Je länger die CPU ausgeschaltet ist, desto weniger Energie wird verbraucht. Kurze Ausführungszeiträume vergrößern den Zeitraum, in dem die CPU ausgeschaltet sein kann. Die Ausführungszeit ist die Zeit die benötigt wird, um alle Instruktionen auszuführen [Dym21]. Jede Instruktion bedarf eine bestimmte Anzahl an CPU-Zyklen. Die Zeit pro Zyklus ist abhängig von der Taktrate der CPU.

Die Ausführungszeit eines Entscheidungswaldes setzt sich zusammen aus der Zeit für die

2 ENTSCHEIDUNGSBÄUME

Feature-Extrahierung, der Evaluierung aller im Ensemble enthaltenen Entscheidungsbäume und der Aggregierungsfunktion. Im schlimmsten Fall muss die gesamte Höhe eines Entscheidungsbaumes traversiert werden, um das Ergebnis zu bestimmen. Aus diesem Grund skaliert die Ausführungszeit mit der traversierten Höhe jedes Baumes.

Um die Instruktionen zu minimieren sollten Datentypen verwendet werden, die von der CPU mit höchstens einem Wort dargestellt werden können. Eine 8-Bit CPU würde zum Laden in Register eines 32-Bit Datentypen vier mal so viele Instruktionen benötigen wie bei einem 8-Bit Datentypen. Außerdem sollten Operationen verwendet werden, die durch native Hardware-Operationen abgebildet werden können. Ist dem nicht so, muss diese Operation durch Software ersetzt werden. Dies erfordert mehr Zyklen als eine native Operation in Hardware.

Zu Beachten bei der Minimierung ist, dass Instruktionen unterschiedlich viele Zyklen benötigen und Funktionsaufrufe Overhead erzeugen. Ein Beispiel dafür ist die Optimierung *Function Inlining* [LM99]. Der Aufruf von Funktionen kann einen hohen Overhead durch den Kontextwechsel erzeugen. Aus diesem Grund verringert diese Optimierung die Ausführungszeit, erhöht aber die Programmgröße signifikant. Im Umkehrschluss könnten durch die Verwendung von Funktionen der Nutzen des Programmspeichers verringert werden, Ausführungszeit und Energieverbrauch aber erhöht werden.

2.5.2 Programmgröße

Die Programmgröße ist die Gesamtheit aller Instruktionen die für das Programm benötigt werden [Dym21]. Dabei ist der Anteil für die Entscheidungswälder integral und der Anteil für die peripären Funktionalitäten zu vernachlässigen. Die Programmgröße, die für einen Entscheidungswald benötigt wird, skaliert mit der Waldgröße und Höhe der einzelnen Entscheidungsbäume.

Die Höhe des Entscheidungsbaumes ist die Verzweigungstiefe der verschachtelten Tests. Jeder Test ist ein Vergleich mit einem Schwellenwert. Die Programmgröße für einen Vergleich setzt sich zusammen aus den Operationen um die Operanden in die Register zu laden und die Instruktion um den Vergleich durchzuführen, sowie Abzweiginstruktionen. Wie in Kapitel 2.5.1 sind Instruktionen durch einen passenden Datentypen zu vermeiden.

Ein weiterer Faktor sind die Instruktionen, die zur Rückgabe des Klassifizierungsergebnis benötigt werden. In Kapitel 2.3 wurden verschiedene Möglichkeiten der Rückgabe diskutiert, die relevant bei dem Aggregierungsprozess eines Ensembles ist. Einerseits kann die Rückgabe

2.5 RESSOURCENBEDARF AUF DEM MIKROCONTROLLER

eine Wahrscheinlichkeitsverteilung sein und andererseits eine diskrete Klasse. Bei m möglichen Klassen würde die erste Variante m -mal so viele Instruktion benötigen, wie die zweite Variante, da der Rückgabektor zuvor mit der Wahrscheinlichkeitsverteilung gefüllt werden muss. In der Praxis werden aber weniger Instruktion benötigt, da es eine große Überschneidung der Wahrscheinlichkeitsverteilungen gibt, die zurück gegeben werden. Die Instruktionen, um den Rückgabektor zu befüllen, können durch *Basic Blocks*, d. h. beschriftete Instruktionsblöcke, geschickt recycled werden. Zudem können Zuweisungen ausgelassen werden, die die Wahrscheinlichkeit 0 zuweisen, da der Vektor mit Nullen initialisiert wird. Dennoch werden signifikant mehr Instruktionen benötigt als bei der diskreten Variante. Aus diesem Grund wurde ein hybrider Ansatz vorgeschlagen, der im Falle eines eindeutigen Ergebnisses mit einer Toleranz von $\epsilon \in [0, 1]$ die diskrete Klasse statt der Wahrscheinlichkeitsverteilung zurück gibt.

2 ENTSCHEIDUNGSBÄUME

Künstliche Neuronale Netze

Das mathematische Modell von künstlichen neuronalen Netzen wurde von McCulloch und Pitts im Jahre 1943 erfunden [MP43]. Dieses Modell ist eine Abstraktion des biologischen Neuronen als logischer Mechanismus.

Das Nervensystem besteht aus einem Netz von Neuronen, die miteinander verbunden sind und über elektrische Impulse miteinander interagieren [Ros61]. Man unterscheidet beim biologischen Neuronen zwischen *Afferent*-Neuronen, *Efferent*-Neuronen und *Inter*-Neuronen. Afferent-Neuronen nehmen elektrische Signale von Organen entgegen und können als *Input* interpretiert werden. Efferent-Neuronen geben elektrische Signale an *Effektorzellen* weiter und können als *Output* interpretiert werden. Inter-Neuronen nehmen elektrische Signale von Afferent-Neuronen oder Inter-Neuronen entgegen und geben sie an Inter-Neuronen oder Efferent-Neuronen weiter. Wenn der Schwellenwert eines *Dendrite* von einem Neuronen durch ein elektrisches Signal erreicht wurde, wird ein elektrisches Signal über den *Axon* an ein anderes Neuron oder Effektorzellen übertragen.

Diese Charakteristiken werden mathematisch als ein Vergleich von einer gewichtete Summe von eingehenden Signalen mit einem Schwellenwert modelliert [HH19]. Gleichung 3.1 stellt diesen Zusammenhang dar.

$$y = \sigma\left(\sum_{i=1}^n \mathbf{w}_i \mathbf{x}_i + b\right) \quad (3.1)$$

Die Vergleichsoperation ist die *Aktivierungsfunktion* $\sigma : \mathbb{R} \mapsto \mathbb{R}$, die in diesem Fall die Stufenfunktion ist. Die Eingabe $\mathbf{x} \in \mathbb{R}^n$ wird mit $\mathbf{w} \in [0, 1]^n$ gewichtet und der *Bias* $b \in \mathbb{R}$ wird addiert. Der Bias stellt den Schwellenwert dar.

Das künstliche neuronale Netz approximiert eine arbiträre Funktion f^* . Dazu findet es eine Menge von Parametern θ , wodurch $f^*(\mathbf{x}) \approx f(\mathbf{x}, \theta)$ möglichst gut von der Approximations-

3 KÜNSTLICHE NEURONALE NETZE

funktion f abgebildet wird [BGC17].

Das KNN ist in Schichten organisiert. Analog zu den biologischen Neuron, gibt es eine *Eingabeschicht* (engl. *Input-Layer*), *Ausgabeschicht* (engl. *Output-Layer*) und *verdeckte Schicht* (engl. *Hidden-Layer*). Dies wird in Abbildung 3.1 illustriert. Analog zur Aktivierung eines einzelnen Neuronen, dargestellt in Gleichung 3.1, stellt Gleichung 3.2 die Aktivierung einer Schicht dar [HH19].

$$\mathbf{a}_l = \sigma_l(\mathbf{z}_l), \quad \mathbf{z}_l := \mathbf{W}_l \mathbf{a}_{l-1} + \mathbf{b}_l \quad (3.2)$$



■ **Abbildung 3.1:** Beispiel eines FFNN mit vier Schichten und einer binären Ausgabeschicht.

Das allgemeine KNN verfügt über $L \in \mathbb{N}$ Schichten. Jede Schicht l verfügt über $n_l \in \mathbb{N}$ Neuronen. Die Aktivierungsfunktion $\sigma_l : \mathbb{R}^{n_l} \mapsto \mathbb{R}^{n_l}$ berechnet die Aktivierung mit der gewichteten Summe \mathbf{z}_l . Die gewichtete Summe setzt sich zusammen aus der Aktivierung der vorherigen Schicht \mathbf{a}_{l-1} die mit $\mathbf{W}_l \in \mathbb{R}^{n_l \times n_{l-1}}$ gewichtet wird. Die Schwellenwerte der Neuronen werden durch die Biase $\mathbf{b}_l \in \mathbb{R}^{n_l}$ dargestellt. Gleichung 3.3 zeigt, wie das allgemeine KNN mit einer rekursiven Funktion modelliert werden kann.

$$\mathbf{a}_1 := \mathbf{x}, \quad \mathbf{z}_l := \mathbf{W}_l \mathbf{a}_{l-1} + \mathbf{b}_l, \quad \mathbf{a}_l := \sigma_l(\mathbf{z}_l), \quad \mathbf{f}(\mathbf{x}) := \mathbf{a}_L \quad (3.3)$$

Diese Arbeit nutzt ausschließlich *Feed Forward neuronale Netzwerke*. Diese werden durch *dichte Schichten* (engl. *Dense-Layer*) charakterisiert, d. h. Schichten in denen alle Neuronen einer Schicht mit allen Neuronen der folgenden Schicht verbunden sind [BGC17].

3.1 Keras

Keras ist die am meisten genutzte *deep learning* API und wurde in Python geschrieben [ker21]. Dadurch ist sie kompatibel mit allen gängigen Betriebssystemen. Ihr Fokus ist eine intuitive und simple API anzubieten, sodass schnelle Iterationen im Entwicklungsprozess möglich sind. Trotzdem ist sie effizient und skalierbar, um die Kapazitäten großer Rechenverbunde auszunutzen.

Keras abstrahiert das ML System *Tensorflow*. Tensorflow implementiert ML Algorithmen, die dem Stand der Forschung entsprechen. Der Fokus ist auf effizientes Training der Modelle [ABC⁺16] gerichtet. Dafür nutzt es die Multikernarchitektur von CPUs, GPUs und spezialisierter Hardware, sogenannten TPUs (**T**ensor **P**rocessing **U**nit), aus. Es wurde als open-source Projekt veröffentlicht und ist weit verbreitet.

Keras bietet die in dieser Arbeit benötigten Algorithmen an, weshalb es zum Trainieren von FFNNs verwendet wird.

3.2 Training von KNN

Um die Zielfunktion f^* zu approximieren ist eine Kostenfunktion nötig, die den Abstand von der approximierten Funktion zur approximierenden Funktion angibt [Nie]. Im Optimierungsprozess wird die Kostenfunktion minimiert. Oft wird die Kostenfunktion auch als *Verlustfunktion* (engl. *loss function*), *Fehlerfunktion* (engl. *error function*), oder *Zielfunktion* (engl. *objective function*) bezeichnet [BGC17]. Je nach Publikation können einiger dieser Funktionen spezielle Bedeutungen haben, in dieser Arbeit haben sie aber die gleiche Bedeutung.

Das KNN wird für eine endliche Anzahl an Trainingsdurchläufen trainiert [Nie]. Ein Trainingsdurchlauf einer Trainingsmenge wird als Epoche bezeichnet. Als *Backpropagation* wird der Algorithmus bezeichnet, der mit Hilfe des Lernalgorithmus den Fehler der Kostenfunktion rückwärts durch die Schichten des KNN propagiert, wodurch die Parameter θ angepasst werden. Der Lernalgorithmus steuert dabei, wie in dieser Epoche die Parameter mit den Trainingsdaten aktualisiert werden. Zu den Parametern gehört die Struktur des neuronalen Netzwerks, Aktivierungsfunktionen, Gewichte und Biase.

Abhängig von der Strategie des Lernalgorithmuses müssen die Gradienten rückwärts, rekursiv Schicht für Schicht zurück propagiert werden. Ziel ist es den Fehler unter Berücksichtigung der Parameter zurück zu propagieren. Dafür ist es nötig die Ableitungen von der Kostenfunktion

zu den Parametern zu berechnen. Gleichung 3.3 zeigt, dass das Ergebnis einer Schicht die Aktivierung der gewichteten Summe des Ergebnisses der jeweiligen vorherigen Schicht ist. Um die Ableitungen unter Berücksichtigung der Parameter zu berechnen ist die Anwendung der Kettenregel nötig.

Anfangen mit dem Gradienten der Aktivierungen der Ausgabeschicht, wird der Fehler rekursiv zurück propagiert. Dabei wird in jeder Schicht die Korrektur auf die Parameter addiert. Anschließend kann der Vorgang für nächste Epoche wiederholt werden.

3.3 Lernalgorithmen

Die Lernalgorithmen bestimmen die Strategie, mit der die Parameter θ des KNN aktualisiert werden [HH19]. In dieser Arbeit wird ADAM verwendet, da dieser Algorithmus die Vorteile von *SGD* (Stochastic Gradient Descent) mit Momentum und *RMSprop* vereinigt [KB14, HH19].

SGD ist eine Approximation von *Gradient Descent* (GD) [BGC17]. GD ist ein iterativer Algorithmus, der den Gradienten in Richtung des Extrema folgt und dementsprechend die Eingabeparameter aktualisiert.

$$\theta_{k+1} := \theta_k - \text{sign}(C'(\theta_k))\eta \quad (3.4)$$

Gleichung 3.4 illustriert diesen iterativen Prozess für Minimierung im eindimensionalen Fall, wobei $\eta > 0$ eine angemessene *Lernrate* ist, θ der Eingabeparameter und C die Kostenfunktion. Ist die Lernrate zu groß könnte keine Verbesserung beobachtet werden, da das Maxima immer übersprungen wird. Ist die Lernrate zu klein könnte die Konvergenz sehr langsam sein. Wenn ein Sattelpunkt erreicht wird ist der Gradient 0. Aus diesem Grund können mit dieser Technik globale Extrema verfehlt werden.

Im mehrdimensionalen Fall wird für jede Komponente des Eingabevektors dieser Prozess durchgeführt, sodass für jede Komponente die Richtung des Extrema verfolgt wird. Dies impliziert, dass GD sehr aufwendig zu berechnen ist für Eingabevektoren mit hohen Dimensionen. Gleichung 3.5 zeigt die iterative Berechnung im mehrdimensionalen Fall.

$$\theta_{k+1} = \theta_k - \nabla C(\theta_k)\eta \quad (3.5)$$

Diese Methode konvergiert, wenn alle Komponenten des Gradienten 0 sind. Je größer die Dimension des Eingabevektor ist, desto aufwendiger ist die Berechnung.

SGD ist eine Approximation von GD. Es nutzt eine zufällige Teilmenge des Eingabevektors, der sogenannte *Mini-Batch*. Es wird angenommen, dass der Gradient des Mini-Batches ähnlich ist zu dem Gradienten des gesamten Eingabevektors. Folglich werden mit jedem Mini-Batch die Parameter des KNN aktualisiert, wodurch in jeder Epoche die Parameter mehr als einmal aktualisiert werden. Ziel dieser Approximierung ist eine Laufzeitverbesserung.

(S)GD mit Momentum versucht zu vermeiden, dass lokale Extrema gefunden werden anstatt globale Extrema, indem Momentum aus vorherigen Gradienten beibehalten wird, um aus lokalen Extrema wieder raus zu finden [HH19]. Gleichung 3.6 zeigt die iterative Berechnung.

$$\mathbf{v}_{-1} = \mathbf{0}, \quad \mathbf{v}_k = \mathbf{v}_{k-1}\gamma + \nabla C(\boldsymbol{\theta}_k), \quad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathbf{v}_k\eta \quad (3.6)$$

Zur Berechnung wird ein Hilfsvektor \mathbf{v} verwendet, welcher das Momentum vergangener Gradienten darstellt. In jeder Iteration fließt ein Anteil γ , typischerweise $\gamma = 0.9$, von dem Hilfsvektor in die Berechnung der neuen Eingabeparameter ein. Der Unterschied zu GD (3.5) ist der Anteil vergangener Gradienten.

RMSprop ist eine Generalisierung von *Adagrad* [MH17]. Adagrad passt die Lernrate η an, sodass Komponenten des Eingabevektors die überrepräsentiert sind eine geringere Lernrate erhalten und unterrepräsentierte Komponenten eine im Verhältnis größere Lernrate [DHS11]. Gleichung 3.7 zeigt, wie sich iterativ die Lernrate antiproportional zur kumulierten Norm der Gradienten der Kostenfunktion verringert [LF19, KB14]. Dabei wird für ϵ eine kleine Zahl gewählt, um Teilen durch 0 zu vermeiden aber keinen signifikanten Einfluss auf die Berechnung zu haben.

$$\mathbf{g}_k = \nabla C_{j_k}(\boldsymbol{\theta}_k), \quad \mathbf{w}_k = \mathbf{w}_{k-1} + \mathbf{g}_k^2, \quad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathbf{g}_k \circ \frac{\eta}{\sqrt{\mathbf{w}_k + \epsilon}} \quad (3.7)$$

Das Problem an Adagrad ist, dass die Lernrate zu schnell gegen 0 konvergieren kann, wodurch das Zielextrema nicht erreicht wird [BGC17]. RMSprop [HSS12] (3.8) löst dieses Problem, indem anstatt die Gradienten zu summieren, ein exponentiell gewichteten gleitenden Mittelwert verwendet wird [BGC17].

$$\mathbf{w}_{-1} = \mathbf{0}, \quad \mathbf{g}_k = \nabla C_{j_k}(\boldsymbol{\theta}_k) \\ \mathbf{w}_k = \mathbf{w}_{k-1}\gamma + \mathbf{g}_k^2(1 - \gamma), \quad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathbf{g}_k \circ \frac{\eta}{\sqrt{\mathbf{w}_k + \epsilon}} \quad (3.8)$$

Gleichung 3.9 zeigt, wie Adam RMSprop und SGD mit Momentum vereint, wobei $\gamma_1 < \gamma_2 < 1$ [KB14]. Adam beinhaltet RMSprop und hat eine Momentum komponente, indem

der Moment der ersten Ordnung mit exponentieller Gewichtung approximiert wird [BGC17]. Zudem korrigiert Adam den Bias der Momente der ersten und zweiten Ordnung der durch die Initialisierung entsteht.

$$\begin{aligned}
\mathbf{v}_{-1} &= \mathbf{w}_{-1} = \mathbf{0}, & \mathbf{g}_k &= \nabla C_{j_k}(\boldsymbol{\theta}_k) \\
\mathbf{v}_k &= (\mathbf{v}_{k-1}\gamma_1 + \mathbf{g}_k(1 - \gamma_1)) / (1 - \gamma_1^k) \\
\mathbf{w}_k &= (\mathbf{w}_{k-1}\gamma_2 + \mathbf{g}_k^2(1 - \gamma_2)) / (1 - \gamma_2^k) \\
\boldsymbol{\theta}_{k+1} &= \boldsymbol{\theta}_k - \mathbf{v}_k \circ \frac{\eta}{\sqrt{\mathbf{w}_k + \epsilon}}
\end{aligned} \tag{3.9}$$

3.4 Aktivierungsfunktionen

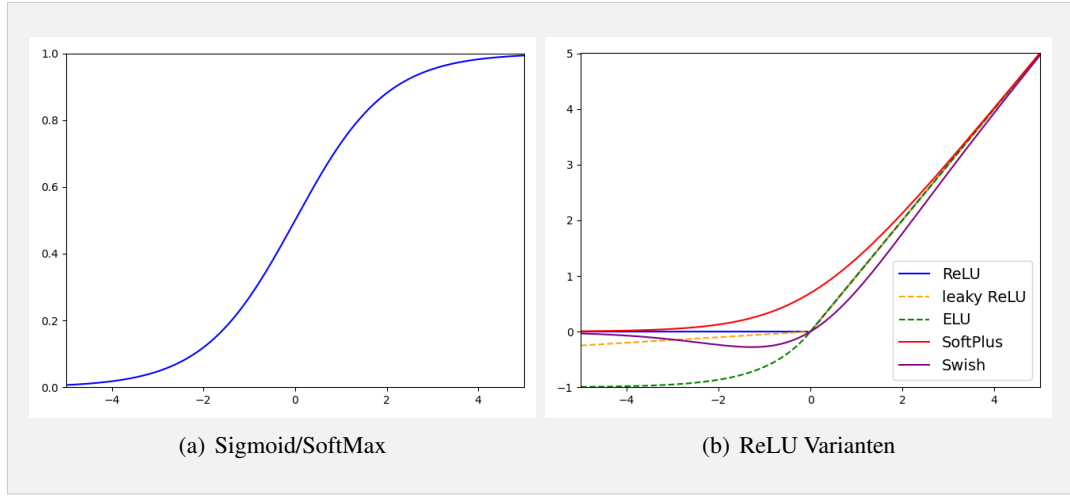
Die Aktivierungsfunktion entscheidet ob ein Neuron aktiviert wird oder nicht [NIGM18]. Sie können entweder linear oder nicht-linear sein. Es ist aber nötig nicht-lineare Funktionen zu verwenden, damit jede kontinuierliche Funktion approximiert werden kann [ADIP21]. Sie unterscheiden sich in ihren Eigenschaften und Berechnungskosten, was eine besondere Rolle für Mikrocontroller spielt.

In der frühen Geschichte der neuronalen Netzwerke wurde die *Sigmoid*-Funktion (3.10) viel verwendet, da sie asymptotisch begrenzt, kontinuierlich und nicht-linear ist [ADIP21]. Oft wird sie heute in der Ausgabeschicht für binäre Klassifizierungsprobleme eingesetzt [NIGM18]. Allerdings ist sie für tiefe neuronale Netzwerke ungeeignet, da der Gradient zwischen 0 und 0.25 ist und dadurch im Backpropagation-Prozess bereits nach wenigen Schichten gegen 0 geht. Die *SoftMax*-Funktion (3.11) oder normalisierte Exponentialfunktion berechnet für einen Eingabevektor eine Wahrscheinlichkeitsverteilung. Die Einträge dieser Verteilung können für die Wahrscheinlichkeiten der einzelnen Klassen eines multivariat Klassifizierungsproblem interpretiert werden.

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{3.10}$$

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3.11}$$

Zur *ReLU*-Familie [ADIP21] gehört die ReLU-Funktion [GBB11, KMK14, EUD18, Alc18]



■ **Abbildung 3.2:** Verschiedene Aktivierungsfunktionen.

und ihre Varianten [MHN13], sowie die *SoftPlus*- [DBB⁺01] und *Swish*-Funktion [RZL17]. ReLU (3.12) steht für „*rectified linear function*“ und wird häufig in modernen neuronalen Netzwerken verwendet [ADIP21]. Sie ist nicht differenzierbar bei 0 und die Ableitung für negative Eingaben ist 0. Dies kann zu *sterbenden Neuronen* (engl. *dying neurons*) führen, da der Bias so negativ wird, sodass das Neuron nicht mehr aktiviert wird. Zudem ist der Trainingsprozess verlangsamt, wenn der Gradient konstant 0 ist. Dafür ist die Ableitung der Funktion ansonsten 1, was den Backpropagation-Prozess vereinfacht, da der Gradient neutral zur Aktivierungsfunktion ist.

$$\text{ReLU}(x) = \max(x, 0) \quad (3.12)$$

Varianten sind beispielsweise *leaky ReLU* [MHN13] (3.13) und *ELU* (*exponential linear unit*) (3.14) [CUH15], welche versuchen die Defizite des konstanten 0 Gradienten zu lösen, indem der negative Teil der Funktion nicht 0 ist.

$$\text{leaky_ReLU}_\alpha(x) = \begin{cases} x & , \text{ wenn } x \geq 0 \\ \alpha x & , \text{ ansonsten } (\alpha \in \mathbb{R}_0^+) \end{cases} \quad (3.13)$$

$$\text{ELU}_\alpha(x) = \begin{cases} x & , \text{ wenn } x \geq 0 \\ \alpha(e^x - 1) & , \text{ ansonsten } (\alpha \in \mathbb{R}_0^+) \end{cases} \quad (3.14)$$

SoftPlus (3.15) ist analytisch, dafür aber aufwendiger zu berechnen im Vergleich zu ReLU [ADIP21].

$$\text{softplus}(x_i) = \ln(e^x + 1) \quad (3.15)$$

Eine weitere Variante ist *Swish* [RZL17] (3.16). Sie ist analytisch aber nicht monoton. Im Vergleich zu ReLU ist sie aufwendig zu berechnen. Ihre Autoren behaupten aber, dass dadurch bessere Ergebnisse erzielt werden können, ohne andere Parameter zu ändern.

$$\text{swish}(x_i) = \frac{xe^x}{e^x + 1} \quad (3.16)$$

3.5 Ressourcenbedarf auf dem Mikrocontroller

Der Speicherverbrauch eines neuronalen Netzes ist abhängig von der Anzahl der Gewichte und Biase, sowie der Größe des verwendeten Datentypen [Kub19]. Ein FFNN mit 3 Schichten der Größe n_1, n_2, n_3 hat $n_1n_2 + n_2n_3$ Gewichte und $n_2 + n_3$ Biase. Die Gewichte und Biase sind für gewöhnlich Gleitkommazahlen. Dieser kann mit 4 und 8 Byte dargestellt werden. Als Alternative können auch Festkommazahlen mit verwendet werden, die 2 Byte benötigen [Gie20].

Die Ausführungszeit des KNN ist stark abhängig von der Hardware. Das KNN kann sowohl im RAM abgelegt werden oder im Flash-Speicher [Eng18]. Wenn es im Flash-Speicher abgelegt werden muss, müssen die Gewichte bei der Ausführung in den RAM und die Register geladen werden. In Yao's Experimenten hat dies die Ausführungszeit um bis zu 74% verlangsamt [Jia17].

Außerdem wird Multiplikation und Division zur Evaluierung des KNNs benötigt [Eng18]. Wenn die Hardware keine Unterstützung dafür anbietet, müssen diese Operationen durch Software ergänzt werden. Diese sind dann signifikant komplexer zu berechnen im Vergleich zu hardwareunterstützten Operationen.

Die Ausführungszeit ist auch abhängig von der Struktur des Netzwerkes [Gie20]. Je mehr Gewichte eine KNN hat, desto mehr Multiplikationen müssen durchgeführt werden. Zudem wird in jeder Schicht eine Aktivierungsfunktion angewendet, die ebenfalls sehr aufwendige Berechnungen benötigen kann [VKK⁺20].

3.5 RESSOURCENBEDARF AUF DEM MIKROCONTROLLER

Es gibt verschiedene Optimierungen, die die Ausführungszeit und Speicherbedarf verringern. Giese hat in seiner Arbeit festgestellt, dass das Löschen von Gewichten (engl. pruning) oder das Gruppieren von ähnlichen Gewichten (engl. quantization) signifikante Verbesserungen sowohl des Speicherverbrauchs und der Ausführungszeit zeigt [Gie20]. Denkbar sind auch Compiler-Optimierungen, die sich sowohl auf Ausführungszeit, Speicherbedarf und Energiebedarf auswirken können. Giese stellte fest, dass sowohl der Speicherbedarf verringert werden kann, als auch die Ausführungszeit.

3 KÜNSTLICHE NEURONALE NETZE

Standortbestimmung

- Location Awareness
- Anwendungen?

4.1 Lokalisierung

- Indoor
- Outdoor
- Network vs. Device based

4.2 Anwendung von Lokalisation

- GPS
- Triangulierung
- Navigation, Monitoring, Tracking

4.3 Standortbestimmung mit maschinellen Lernen

- Vision Based

4.4 Orientierungssinn

- Beispiel Honigbienen

4 STANDORTBESTIMMUNG

Trainings- und Validationsdaten

- Daten aufgenommen mit CoppeliaSim(?)

5.1 Simulierten Sensordaten

- Wie wurde es aufgenommen?
- Welche Sensoren
- Aufgenommene Routen

5.2 Künstlichen Sensordaten

- Motivation: Warum ist das nötig?

5.2.1 Magnetfeld

- Welchen Sensor spiegelt das wieder?
- Wie funktioniert das Modell?
- Was und Wie wurden Daten ergänzt?

5.2.2 Temperatur

- Welchen Sensor spiegelt das wieder?
- Wie funktioniert das Modell?
- Was und Wie wurden Daten ergänzt?

5.2.3 Lautstärke

- Welchen Sensor spiegelt das wieder?
- Wie funktioniert das Modell?
- Was und Wie wurden Daten ergänzt?

5.2.4 WLAN Zugangspunkte

- Welchen Sensor spiegelt das wieder?
- Wie funktioniert das Modell?
- Was und Wie wurden Daten ergänzt?

5.3 Simulation von Interrupts

- Motivation => Energieverbrauch, Spiegelung der echten Datenaufnahme, Reduzierung der Trainingsdaten
- Wie funktioniert ist?
- Wie und Wann bei der Datenverarbeitung wird es gemacht?

5.4 Standortenkodierung

- Wie werden Orte modelliert?
- Warum wurden die Orte so modelliert?
- Wie werden Pfade zwischen Orten modelliert?
- Sollte das Modell ausgeben können, dass kein Ort erkannt wurde?
- Sollten Pfade erkannt werden können?

5.5 Feature-Extrahierung

- Datenfenster: Realzeit vs. Diskret mittels Wakeups
- Relevanz von Zeit
- Welche Feature werden genutzt?

- Wie werden diese extrahiert?
- Welchen Mehrwert verschaffen diese Features?
- Welchen Einfluss haben Sie im Hinblick auf Ressourcenbedarf(?), Klassifizierungsgenauigkeit(?), Fehlertoleranz(?)

5.6 Aufteilung der Daten

- Kurz und knapp wie und warum werden die Daten aufgeteilt.
- Sollten Trainingsdaten um synthetische Daten ergänzt werden?
 - ◆ Fault Daten, um das Modell Robuster zu machen
 - ◆ Synthetische Routen
- Wie viele Trainingsdaten werden benötigt?
 - ◆ Um KNN zu trainieren?
 - ◆ Um Entscheidungsbaum zu trainieren?
 - ◆ Ggf. Unterschiede klären

5 TRAININGS- UND VALIDATIONSDATEN

ML-Modelle

TODO

6.1 Entscheidungswald

- Wie wird automatisch das beste Modell gefunden?
- Warum trainieren wir es so => Um die beste Konfiguration zu finden unter den bestehenden Restriktionen.
- Welche Restriktionen werden eingefordert?
- Warum gibt es diese Restriktionen?
- Welchen Einfluss hätte es, gäbe es diese nicht?

6.2 Feed Forward Neuronales Netzwerk

- Wie viele Hidden Layers und Neuronen sollte es haben?
- Welche Aktivierungsfunktionen werden verwendet?
- Braucht man mehr Neuronen/Hidden Layer mit steigender Ort Anzahl?

6.3 Feedback Kanten

- Was ist das?
- Wofür ist das Relevant?
- Wie funktionier es?

6.4 Training der Modelle

- Trainieren in Phasen => Erklären wie es genau funktioniert?
- Warum trainieren wir in Phasen?
- Sag das wir die Zyklen der einzelnen Pfade nutzen
- Werden mehr Trainingsdaten benötigt mit steigender Ort Anzahl? Wenn ja wie viel?

Evaluation

TODO

7.1 Klassifizierungsgenauigkeit

- Metriken
- Wie groß ist die Wahrscheinlichkeit, dass die Orte korrekt erkannt werden?
- Entscheidungsbaum vs KNN
- Skalierung mit Anzahl der Orte
- Signifikanz der Features
- Einfluss von einzelnen Features für die Klassifizierungsgenauigkeit(?)
- Ist es sinnvoll für jeden Ort ein Feature zu haben, dass auf eins gesetzt wird, wenn der Ort erkannt wurde und ansonsten exponentiell abfällt. Wie schnell sollte es fallen, wenn ja?

7.2 Fehlertoleranz

- Metriken => Insbesondere continued_predict hier erwähnen
- Wenn falscher Ort erkannt wurde, wie lange dauert es um wieder den korrekten Ort zu finden?
- Was passiert wenn Sensoren ausfallen?
- Transportbox nicht dem trainierten Pfad folgt?

7 EVALUATION

- Änderungen der Fabrik, e.g. Licht, Wärme, Magnet, Schnelligkeit der Fließbänder
- Einfluss von einzelnen Features für die Fehlertoleranz(?)

7.3 Ressourcennutzung

- Metriken(?)
- Entscheidungsbaum Ausführung
- FFNN Ausführung
- Feature extrahierung
 - ◆ Verhältnis von Kosten zu Nutzen
- Daten Sammlung => Interrupts (Wakeups)
- Einfluss von einzelnen Features für den Ressourcenverbrauch(?)

Diskussion

TODO

8 DISCUSSION

Schlussfolgerungen

TODO

- Kurze, knappe und gut formulierte Schlussfolgerung
 - ◆ Was ist besser Entscheidungsbaum oder Entscheidungswald? Welche Vor- und Nachteile?
 - ◆ Sollten Entscheidungswälder verwendet werden?
 - ◆ Sollten KNNs verwendet werden?
 - ◆ Is die Realzeit relevant für die Ortung(?)
- Kurze Zusammenfassung wichtigster Dinge
- Zukünftige Arbeit

9 SCHLUSSFOLGERUNGEN

Inhalt des USB-Sticks

A INHALT DES USB-STICKS

Literaturverzeichnis

- [ABC⁺16] ABADI, Martín ; BARHAM, Paul ; CHEN, Jianmin ; CHEN, Zhifeng ; DAVIS, Andy ; DEAN, Jeffrey ; DEVIN, Matthieu ; GHEMAWAT, Sanjay ; IRVING, Geoffrey ; ISARD, Michael u. a.: Tensorflow: A system for large-scale machine learning. In: *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, S. 265–283
- [ADIP21] APICELLA, Andrea ; DONNARUMMA, Francesco ; ISGRÒ, Francesco ; PREVETE, Roberto: A survey on modern trainable activation functions. In: *Neural Networks* (2021)
- [Alc18] ALCAIDE, Eric: E-swish: Adjusting activations to different network depths. In: *arXiv preprint arXiv:1801.07145* (2018)
- [BFSO84] BREIMAN, Leo ; FRIEDMAN, Jerome ; STONE, Charles J. ; OLSHEN, Richard A.: *Classification and regression trees*. CRC press, 1984
- [BGC17] BENGIO, Yoshua ; GOODFELLOW, Ian ; COURVILLE, Aaron: *Deep learning*. Bd. 1. MIT press Massachusetts, USA:, 2017
- [Bre96] BREIMAN, Leo: Bagging predictors. In: *Machine learning* 24 (1996), Nr. 2, S. 123–140
- [Bre01] BREIMAN, Leo: Random forests. In: *Machine learning* 45 (2001), Nr. 1, S. 5–32
- [CUH15] CLEVERT, Djork-Arné ; UNTERTHINER, Thomas ; HOCHREITER, Sepp: Fast and accurate deep network learning by exponential linear units (elus). In: *arXiv preprint arXiv:1511.07289* (2015)
- [D⁺02] DIETTERICH, Thomas G. u. a.: Ensemble learning. In: *The handbook of brain theory and neural networks* 2 (2002), S. 110–125
- [DBB⁺01] DUGAS, Charles ; BENGIO, Yoshua ; BÉLISLE, François ; NADEAU, Claude ; GARCIA, René: Incorporating second-order functional knowledge for better option pricing. In: *Advances in neural information processing systems* (2001), S. 472–478
- [DHS11] DUCHI, John ; HAZAN, Elad ; SINGER, Yoram: Adaptive subgradient methods for online learning and stochastic optimization. In: *Journal of machine learning research* 12 (2011), Nr. 7
- [Dym21] DYMEL, Tom: Handgestenerkennung mit Entscheidungsbäumen. (2021), 02, S. 1–71
- [Efr92] EFRON, Bradley: Bootstrap methods: another look at the jackknife. In: *Breakthroughs in statistics*. Springer, 1992, S. 569–593
- [Eng18] ENGELHARDT, Sebastian: *Optische Gestenerkennung mit künstlichen neuronalen Netzen für kleine eingebettete Systeme*. 10 2018
- [Ent20a] ENTWICKLER scikit-learn: *1.10.6. Tree algorithms: ID3, C4.5, C5.0 and CART*. <https://scikit-learn.org/stable/modules/tree.html#tree-algorithms>. Version: 2020

LITERATURVERZEICHNIS

- [Ent20b] ENTWICKLER scikit-learn: *1.11. Ensemble methods*. <https://scikit-learn.org/stable/modules/ensemble.html#ensemble>. Version: 2020
- [EUD18] ELFWING, Stefan ; UCHIBE, Eiji ; DOYA, Kenji: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. In: *Neural Networks* 107 (2018), S. 3–11
- [FS97] FREUND, Yoav ; SCHAPIRE, Robert E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: *Journal of computer and system sciences* 55 (1997), Nr. 1, S. 119–139
- [GBB11] GLOROT, Xavier ; BORDES, Antoine ; BENGIO, Yoshua: Deep sparse rectifier neural networks. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics JMLR Workshop and Conference Proceedings*, 2011, S. 315–323
- [GEW06] GEURTS, Pierre ; ERNST, Damien ; WEHENKEL, Louis: Extremely randomized trees. In: *Machine learning* 63 (2006), Nr. 1, S. 3–42
- [Gie20] GIESE, Anton: *Compression of Artificial Neural Networks for Hand Gesture Recognition*. 10 2020
- [GL09] GHOSH, Joydeep ; LIU, Alexander: K-Means. In: *The top ten algorithms in data mining* 9 (2009), S. 21–22
- [HH19] HIGHAM, Catherine F. ; HIGHAM, Desmond J.: Deep learning: An introduction for applied mathematicians. In: *SIAM Review* 61 (2019), Nr. 4, S. 860–891
- [HSS12] HINTON, Geoffrey ; SRIVASTAVA, Nitish ; SWERSKY, Kevin: Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. (2012)
- [Jia17] JIADONG, Yao: *Performance of Artificial Neural Networks on an AVR Microcontroller*. 2017
- [KB14] KINGMA, Diederik P. ; BA, Jimmy: Adam: A method for stochastic optimization. In: *arXiv preprint arXiv:1412.6980* (2014)
- [ker21] *Keras Dokumentation*. <https://keras.io/about/>. Version: 2021
- [KMK14] KONDA, Kishore ; MEMISEVIC, Roland ; KRUEGER, David: Zero-bias autoencoders and the benefits of co-adapting features. In: *arXiv preprint arXiv:1402.3337* (2014)
- [Kub19] KUBIK, Philipp: *Zuverlässige Handgestenerkennung mit künstlichen neuronalen Netzen*. 04 2019
- [LF19] LYDIA, Agnes ; FRANCIS, Sagayaraj: Adagrad—An optimizer for stochastic gradient descent. In: *Int. J. Inf. Comput. Sci.* 6 (2019), Nr. 5
- [LM99] LEUPERS, Rainer ; MARWEDEL, Peter: Function inlining under code size constraints for embedded processors. In: *1999 IEEE/ACM International Conference on Computer-Aided Design. Digest of Technical Papers (Cat. No. 99CH37051)* IEEE, 1999, S. 253–256
- [LR76] LAURENT, Hyafil ; RIVEST, Ronald L.: Constructing optimal binary decision trees is NP-complete. In: *Information processing letters* 5 (1976), Nr. 1, S. 15–17
- [MH17] MUKKAMALA, Mahesh C. ; HEIN, Matthias: Variants of rmsprop and adagrad with logarithmic regret bounds. In: *International Conference on Machine Learning PMLR*, 2017, S. 2545–2553
- [MHN13] MAAS, Andrew L. ; HANNUN, Awni Y. ; NG, Andrew Y.: Rectifier nonlinearities improve neural network acoustic models. In: *Proc. icml* Bd. 30 Citeseer, 2013, S. 3

- [MP43] MCCULLOCH, Warren S. ; PITTS, Walter: A logical calculus of the ideas immanent in nervous activity. In: *The bulletin of mathematical biophysics* 5 (1943), Nr. 4, S. 115–133
- [Nie] NIELSEN, Michael: Neural Networks and Deep Learning.
- [NIGM18] NWANKPA, Chigozie ; IJOMAH, Winifred ; GACHAGAN, Anthony ; MARSHALL, Stephen: Activation functions: Comparison of trends in practice and research for deep learning. In: *arXiv preprint arXiv:1811.03378* (2018)
- [PGP98] PEI, M ; GOODMAN, ED ; PUNCH, WF: Feature extraction using genetic algorithms. In: *Proceedings of the 1st International Symposium on Intelligent Data Engineering and Learning, IDEAL* Bd. 98, 1998, S. 371–384
- [PVG⁺11] PEDREGOSA, F. ; VAROQUAUX, G. ; GRAMFORT, A. ; MICHEL, V. ; THIRION, B. ; GRISEL, O. ; BLONDEL, M. ; PRETTENHOFER, P. ; WEISS, R. ; DUBOURG, V. ; VANDERPLAS, J. ; PASSOS, A. ; COURNAPEAU, D. ; BRUCHER, M. ; PERROT, M. ; DUCHESNAY, E.: Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830
- [Qui86] QUINLAN, J. R.: Induction of decision trees. In: *Machine learning* 1 (1986), Nr. 1, S. 81–106
- [Qui90] QUINLAN, J R.: Decision trees and decision-making. In: *IEEE Transactions on Systems, Man, and Cybernetics* 20 (1990), Nr. 2, S. 339–346
- [Qui14] QUINLAN, J R.: *C4. 5: programs for machine learning*. Elsevier, 2014
- [Ros61] ROSENBLATT, Frank: Principles of neurodynamics. perceptrons and the theory of brain mechanisms / Cornell Aeronautical Lab Inc Buffalo NY. 1961. – Forschungsbericht
- [RZL17] RAMACHANDRAN, Prajit ; ZOPH, Barret ; LE, Quoc V.: Searching for activation functions. In: *arXiv preprint arXiv:1710.05941* (2017)
- [Ste09] STEINBERG, Dan: CART: classification and regression trees. In: *The top ten algorithms in data mining* 9 (2009), S. 179–201
- [Ven21] VENZKE, Marcus: TODO: Antrag Forschungsprojekt. (2021), 04, S. 1–71
- [VKK⁺20] VENZKE, Marcus ; KLISCH, Daniel ; KUBIK, Philipp ; ALI, Asad ; MISSIER, Jesper D. ; TURAU, Volker: *Artificial Neural Networks for Sensor Data Classification on Small Embedded Systems*. 2020