

Hochschule Reutlingen

Reutlingen University

– Studiengang Mechatronik Bachelor –
Bachelor–Thesis

Entwicklung eines autonomen Systems zur Bilderkennung mithilfe Neuronaler Netze auf dedizierter Hardware

Manuel Barkey
Pestalozzistraße 29
72762 Reutlingen

Matrikelnummer : 762537

Betreuer: Eberhard Binder
Zweitbetreuer: Christian Höfert
Abgabedatum: TT.MM.JJJJ



Inhaltsverzeichnis

1 Einleitung	3
2 Grundlagen	5
2.1 Machine Learning	5
2.2 Deep Learning und Computer Vision	8
2.3 Hardware	10
3 Anforderungen und Analyse	11
3.1 Ziel der Arbeit	11
3.2 Related Work	11
4 Realisierung Objekt Erkennung	13
4.1 Dataset	13
4.2 Training	14
4.3 Parameter Optimierung	15
5 Evaluierung	17
5.1 Evaluierungs Metriken	17
5.2 Ergebnisse	17
6 Entwicklung der Anwendung	21
6.1 Aufbau	21
6.2 OpenVino Toolkit	22
6.3 Raspberry Pi Kamera	24
6.4 Server-Client-Connection	24
6.5 Anwendung gesamt	24
7 Test und Validierung	25
8 Zusammenfassung und Ausblick	27
A Beispiel für ein Kapitel im Anhang	31
A.1 Bsp für ein Abschnitt im Anhang	31

Kapitel 1

Einleitung

Im Rahmen der Bachelor Arbeit wurde ein Überwachungssystem zur Wildtiererkennung, entwickelt, welches auf einem Raspberry Pi läuft und den Nutzer bei Erkennung bestimmter Tiere automatisch benachrichtigt, sowie das Bild an einen Server sendet.

Die Erkennung der Tiere erfolgte mithilfe Neuronaler Netze, wodurch es möglich ist die Überwachung gezielt nur auf bestimmte, relevante Tiere anzuwenden und so den Datenverkehr gering zu halten.

Die Inferenz der Neuronalen Netze wurde dabei auf einer separaten Hardware, dem Neural Compute Stick 2 von Intel ausgeführt.

Des weiteren wurde eine Infrarotfähige Kamera verwendet, damit das System auch in der Nacht einsetzbar ist.

Kapitel 2

Grundlagen

Im folgende Kapitel wird zunächst auf die Grundlegenden Techniken des Machine Learings, insbesondere auf die für die Bilderkennung verwendeten Convolutional Neural Networks eingegangen. Anschließend wird es um die verwendete Hardware, den Neural Compute Stick 2 un seine Anwendungen gehen.

2.1 Machine Learning

Beim Machine Lerning, welches ein Teilgebiet der Computerwissenschaften ist, geht es um Algorithmen, die Zusammenhänge in großen Datenmengen erkennen sollen, ohne explizit darauf programmiert worden zu sein.

Eine Form davon ist das *Supervised Learning*, bei der das Programm neben den Input Daten auch die Zugehörigen Ausgaben erhält und daraus dann die Regeln für Zusammenhänge herleiten soll. Dadurch unterscheidet sich das Vorgehen wesentlich zur klassischen Programmierung, bei der die Regeln vorab definiert werden.



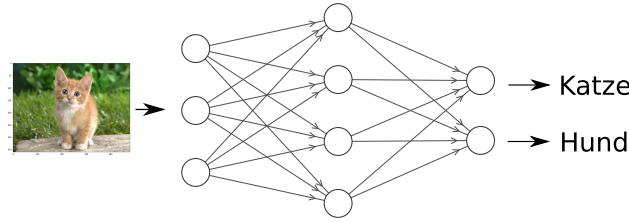
Das herleiten der Regeln erfolgt beim Machine Learning dabei in einem iterativen Prozess, welcher als Training bezeichnet wird. Dabei soll eine math. Funktion, welche die Zusammenhänge beschreibt numerisch angenähert werden. Ist der Zusammenhang linear, spricht man von einer Regression, handelt es sich um Kategorische, liegt ein Klassifizierung problem vor.

Weitere Formen neben dem *Supervised Learning* sind das *Unsupervised Learning*, bei der das Programm keine Labels erhält, sondern diese durch Clustering Verfahren selber finden soll, oder das *Reinforcement Learning*, bei dem das Programm mit der Umwelt interagieren soll.

Da hier jedoch ausschließlich mit dem Supervised Learning gearbeitet wurde, werden diese Techniken nicht näher erläutert.

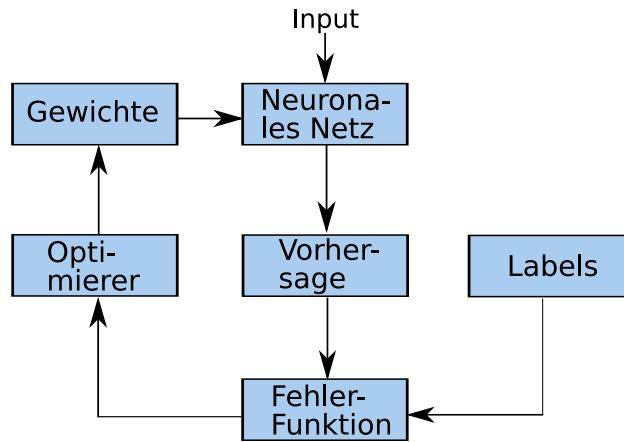
2.1.1 Künstliche Neuronale Netze

Für komplexe Input Daten, wie beispielsweise Bilder, bei denen die einzelnen Pixelwerte als Inputs und der Inhalt des Bildes als Output dienen, werden in der Regel künstliche Neuronale Netze verwendet. Diese sind eine Form des Machine Learings und bestehen aus einer vielzahl an miteinander verbundener Neuronen. Durch unterschiedlich starke Gewichtungen der einzelnen Verbindungen, auch Gewichte genannt, können für unterschiedliche Input Daten die entsprechenden Outputs gefunden werden.



Die richtige einstellung der Gewichte, welche zunächst zufällig initialisiert werden, erfolgt dabei im Trainingsprozess, welcher in ?? schematisch dargestellt ist und sich in die drei Schritte:

- Feed Forward anhand aktueller Gewichte vorhersage aus den Inputs treffen
- Lossfunction Abweichung zu tatsächlichen werten bestimmen
- Backpropagation minimierung der Fhlerfunktion durch anpassung der Gewichte



Durch häufiges wiederholen dieser Schritte kann die Fehlerfunktion soweit minimiert werden, dass das Modell auch für neue Input Daten die richtigen Aussagen treffen kann.

Vorwärts

Im Vorwärtsdurchgang wird der Input durch alle Schichten hindurch gereicht, um in der letzten Schicht den gewünschten Output zu liefern. Dabei erhält jedes Neuron wie in 2.1 dargestellt, die Ausgaben aller neuronen der vorherigen Schicht, summiert diese auf und übergibt den Wert einer Aktivierungsfunktion, die den Wert auf einen bestimmten Bereich Skaliert.

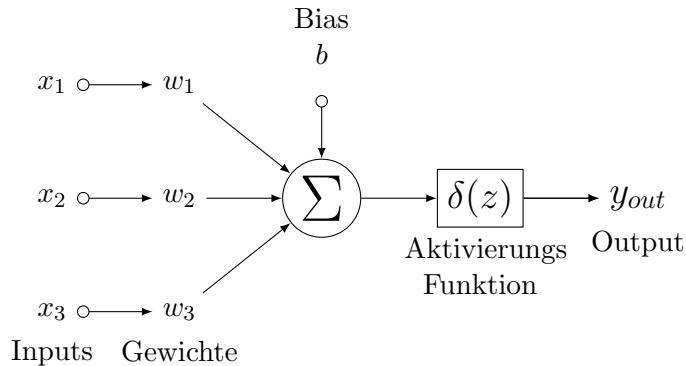


Abbildung 2.1: Einzelnes Perzeptron

Die Berechnung des Vorwärtsdurchgangs von einer ges Schicht zur nächsten, lässt sich die mithilfe der Matrixixmultiplikation durchführen, was Gl. in ... als Vektorschreibweise ergibt.

$$y = a(WTX)$$

wobei $a()$ die Aktivierungsfunktion.
Aktivierungsfunktion können sein

$$\delta(z) = \max(0, z) \quad (2.1)$$

ReLU in den hidden Layer oder

$$\delta(z) = \frac{e^z}{\sum e^x} \quad (2.2)$$

$$\delta(z) = \frac{1}{1 + e^{-x}} \quad (2.3)$$

sigmoid(bin) oder softmax(cat) im letzten layer

bei softmax erhält man Wahrscheinlichkeitsverteilung über allen Output neuronen.

Neben dem Ansatz des Gradienten für die Optimierungs gibt es noch weitere, effizientere verfahren wie z.B. Momentum oder Adam.

Fehlerfunktion

Die Abweichung der Schätzung, welche an den Neuronen der letzten Schicht vorliegen, zu den tatsächlichen Werten, den Labels, wird mithilfe geeigneter Fehlerfunktion bestimmt. Für Regression z.B. abs oder rms und für Kategorisch häufig logarithmisch.

hier am Beispiel einer binären Klassifikation (erg 0 oder 1) mit log loss (crossentropy) dargestellt.

$$L = \hat{y} \log(y) + (1 - \hat{y}) \log(1 - y) \quad (2.4)$$

Durch den Logarithmus wird der Loss um so größer, je weiter die Schätzung y vom tatsächlichen Wert \hat{y} abweicht.

Backpropagation

Durch Berechnung des Gradienten der Fehlerfunktion kann ermittelt werden in welche Richtung die Gewichte angepasst werden müssen, sodass sie sich im nächsten Durchgang minimiert. Dafür wird die Fehlerfunktion für jede Schicht partiell nach den Gewichten abgeleitet, was wie in gl. 2.5 dargestellt mithilfe der Kettenregel für die Aktivierungsfunktion geschieht.

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial w} \quad (2.5)$$

Damit werden die Gewichte dann nach Gleichung 2.6 angepasst.

$$w \leftarrow w - \eta \frac{\partial L}{\partial w} \quad (2.6)$$

wobei die *Learning rate* η die Schrittweite mit der die Anpassung vorgenommen werden soll angibt.

2.1.2 Validierung und Overfitting

um überprüfen zu können ob ein Modell die Trainingsdaten tatsächlich generalisiert hat, dh auch für neue Daten anwendbar ist, oder diese nur auswendig gelernt hat, wird häufig der Datensatz in einen Trainingsanteil und einen Testanteil aufgeteilt.

Mit dem Testdatensatz wird dann schon während des Trainings regelmäßig zwischen geprüft, verengert sich irgendwann nur noch der Fehler der Trainingsdaten, findet Overfitting statt.

Häufig sind zu wenige Trainingsdaten oder zu komplexe/überparametrisierte Modelle und damit zuviele Freiheitsgrade, Grund für Overfitting.

Techniken um Overfitting zu vermeiden sind z.B.

- Augmentierung der Daten

- Regularisierung der Parameter (L1/L2)
- Dropout
- early stopping

Bei Augmentierung werden aus den vorhandenen Daten künstlich mehr Daten generiert, in dem an den Bildern geometrische transformationen oder manipulationen der pixelwerte vorgenommen werden.

Bei Regularisierung wird an die Lossfuction als weiterer Term eine aufsummierung der Gewichte gehängt, wodurch diese bei der Minimierung klein gehalten werden, wodurch weniger potential zur überanpassung da ist.

$$J(w) = E + \lambda \sum_i w_i^2 \quad (2.7)$$

Beim Dropout werden zufällig gewichte zu 0 gesetzt.

early stopping: stoppen des trainings, wenn sich overfitting einstellt.

2.2 Deep Learning und Computer Vision

Das maschinelle sehen verwendet Deep Learining Techniken (CNNs) zusammen mit Techiniken der Digitalen Bildverarbeitung. Die für die Bilderkennung am häufigsten eingesetzte Art der NNs sind CNNs.

Es geht im wesentlichen darum Bilder und videos zu Klassifizieren oder objekte in ihenn zu Lokalisieren.

2.2.1 Convolutional Neural Networks

Um Bilder /Inhalte mithilfe Neuronaler Netze zu erkennen, werden die einzelnen Pixelwerte der Bilder als Inputs verwendet und das auf dem Bild zu erkennende Objekt als output verwendet. Bilder werden als Matrizen der Form $height \times width \times colorchannels$ dargestellt.

Da dies für regulere/vollständig verbundene Neuronale Netze eine enorme Anzahl an Parametern und damit einhergehender rechenkost bedeuten würde, werden hier CNNs verwendet, eine Architektur in der Paramer von verschiedenen Neuronen gemeinsam genutzt werden.

Hauptbestandteil von CNNs sind die *Convolutional Layers* welche die mathematische Faltungsoperation zwischen Input Bild und Filter/Kernel durchführen.

Die Filter meist der Form 3×3 oder 5×5 und mit der selben Tiefe wie der Input, werden während des *Forward Pass/bropagation* zeilenweise über das Bild geschoben und an jeder Stelle das Kreuzprodukt berechnet. Jedes Ergebnis dieser berechnungen ergibt einen Pixel Wert der nächsten Schicht auch Feature Map genannt. ??

Ein weiterer bestandteil von CNNs sind die Pooling Layer, welche eine bestimmte Anzahl an Pixeln zB 3×3 zu einem Wert zusammenfassen wodurch sich die Parameter Anzahl des Bildes verringert. Das hat den Vorteil, dass

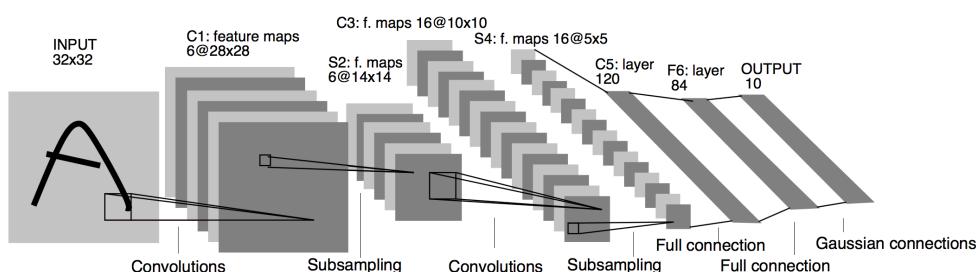


Abbildung 2.2: LeNet-5 cite lecun

Ziel dieser Operation ist es, dass die Filter Maps bestimmte Muster/features die zu einer bestimmten klasse gehören lernen. Mit diesen Filtern können dann unabhängig wo im Bild befindlich, die features wie zB horizontale/vertikale Linien, Ecken oder Kreise gefunden werden.

Beispiel:

$$\begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{pmatrix} \quad (2.8)$$

erkennt vertikale Linien im Bild.

Conv layer:

Faltung an CNN erklärt: input image als (h, w, c) tensor wird mit filter/kernel gefaltet. daraus erhält man feature map zusammen mit pool layer:

pool layer erklärt

ergibt Grundstruktur von CNN

weitere Layer wie dropout

2.2.2 Transfer Learning

erklären, dass Features von einfach bis immer komplexer werdende Muster enthalten, die im Bild zu finden sind.

Filter können zufällig initialisiert und gelernt werden, oder von vorgeübten Netzen wieder verwendet werden. (Transfer Learning oder fine tuning) da die Features (besonders in den vorderen Layern) immer ähnlich sind und das neu Lernen zeitaufwändig und oft sogar ungenauer ist.

je nach Ähnlichkeit des eigenen Datensatzes zu dem auf das Netz ursprünglich trainiert wurde:

scratch, fine tuning, Feature Extractor

2.2.3 Competitions mit Imagenet und co + CNN winner

zuerst competition erklären

dann chronologische Gewinner netz + Besonderheit

2.2.4 Objekt Erkennung

Unterschied deutlich machen: Klassifikator kann nur ein Bild auswerten und wahrscheinlich angeben welche Klasse darauf. Keine Lokalisierung und keine Multi-Objekt.

3 Arten der Bilderkennung: Klassifizierung, Objekt Erkennung (für Multi-Box), Segmentation (jeden Pixel)

dafür Objekt Erkennung notwendig:

Single Shot Detektoren

Two Stage Detektoren

2.2.5 Machine Learning Frameworks

Die Algorithmen müssen nicht jedesmal neu implementiert werden. Für die gängigen Verfahren gibt es Frameworks, welche die Implementierung enthalten und über APIs verwendet werden können (Bsp Tensorflow und Keras)

2.3 Hardware

allg zu hardware für deeplearning. Das besser auf gpu als cpu. weitere: tpu, fpga, vpu, wie zb ncs2.

2.3.1 Neural Compute Stick 2

technischen spezifikationen

2.3.2 AI on the edge

was bedeutet dies. cloud unabhängig und ohne groß rechner. bsp anwendungen.

Kapitel 3

Anforderungen und Analyse

3.1 Ziel der Arbeit

End to end Prozess von Datensatz beschaffung, über training eines geeigneten Neuronalen Netzes bis hin zu implementierung der Applikation, die auf einem Raspberry Pi läuft und die Inferenz auf dem NCS2 ausführt.

Es sollen in Wild Tiere erkannt werden, die in Deutschland heimisch sind.

Das system soll autonom laufen und den Nutzer informieren (und das erkannte bild senden) sobald etwas erkannt wurde.

Im Optimalfall soll es mithilfe Infrarot kamera auch im Dunkeln Tiere erkennen, (da Nachts mehr tiere zu sehen sein werden)

Verwender werden soll: für Inferenz der in 2.3 beschriebene Neural Compute Stick 2, und für die Stuerung der (Einptaininen Computer) RaspberryPi2.

Um auch im Dunklen oder bei nacht Tiere erkennen zu können soll eine Kamera ohne Infrarot Filter verwendet werden. (evtl noch auf realsense eingehen)

Die Kommunikation zwischen Raspberry und Pc soll über eine server/client tcp Verbindung erfolgen. Die Applikattio soll mitteilen wenn etwas erkannt wurde und das bild zusenden. Ausserdem soll das aktuelle frame abgefragt werden können sowie einstellungen bezüglich infrarot leds vorgenommen werden können.

3.2 Related Work

hier:

- Gibt einen Überblick über verwandte Arbeiten im Gebiet
- Strukturiert und Gruppert diese Arbeiten sinnvoll
- Deckt möglichst alle relevanten Arbeiten ab
- Erklärt kurz deren Inhalt und was sie von anderen Arbeiten (vor allem der Eigenen!) abheben
- Positioniert die eigene Arbeit im Gebiet

Kapitel 4

Realisierung Objekt Erkennung

4.1 Dataset

Da es sich um supervised Learning handelt müssen trainings daten gelabelt werden. für validierung und test muss datensatz zu 80, 10, 10 in test/train/validation aufgeteilt werden. wie in 2.1 beschrieben dient das validierungs set zur überwachung während des trainings für overfitting.

Mit dem Test set kann nach dem training die Inferenz also das ausführen des traininierten models getestet werden.

Für die Objekterkennung muss der Datensatz wie in 2.2 beschrieben neben den gelabelten bildern auch die X- und Y-Koordinaten der Bounding Boxen, welche das objekt auf dem Bild umramt, enthalte. das Objekt befindet enthalten.

4.1.1 Datenbeschaffung

Da das Erstellen eines eingangs erwähnten Datensets von Hand sehr müsam ist wird meistens auf Quellen zurückgegriffen, die schon gelabelte Daten zu bestimmten Klassen zur verfüigung stellen. Neben den in 2.2.3 vorgestellten Seiten bietet OpenImages einen vielzahl an Klassen an, darunter auch Unter der Kategorie Säugetiere eine Auswahl an Wild Tier, welche im folgenden verwendet wurde.

Mit einem Open source Tool [?] konnten eine teilmenge aus dem gesammten Open Images datensatzes herunter geladen werden.

Die Label Files haben das anotierungs format `class,xmin,ymin,xmax,ymax` welches wie in 4.1.3 beschrieben wird, noch in ein für tensorflow vertändliches format gebracht werden musste.

Die Verteilung der Klassen im Datensatz war nicht ausbalanciert, wie in ?? zu sehen ist.

Das kann zur folge haben das.

Allg wie viele samples sollte man haben.

Augmentierung im folgenen Teil beschrieben.

4.1.2 Augmentierung

was ist Augmentierung
wie wurde es angewandt
bsp bilder

4.1.3 TF Record Files

was es ist
wie es erstellt wurde

Wie in 4.1.1 erwähnt verwendet tensorflow ein bestimmtes Format für das Datenset, sog Protocol Buffer TF Record Files, Dateien im Binary Format die sowohl die Bilder als auch die Labels enthalten. Das sind Protocol Buffer welche die Daten serialisieren.

Evtl hier besp Ausschnitt von Aufbau eines Proto Elements.

Um nun die von OpenImages heruntergeladenen Bilder und Label Files in das TFRecords Format zu bringen waren mehrere Schritte nötig.

OI - VOC - csv - tf.records

4.2 Training

4.2.1 TF obj det api

Für das Training wurde das Framework Tensorflow verwendet, welches eine API für Objekterkennung bietet.

Welche pretrained Modell gibt es und welche kamen in Frage (für NCS2)? Die Tensorflow Object Detection API bietet eine Vielzahl an vorgebildeten Modellen, dabei wurden die meisten auf den COCO Datensatz trainiert.

Speed/Acc Trade Off

Daneben war bei der Auswahl die Kompatibilität zu OpenVINO zu berücksichtigen: Liste von kompatiblen Modellen.

Trainiert wurde auf:

- SSD MobileNet und Inception
 - keine Region Proposal, dafür vordefinierte Ankerboxen, und CNN mit unterschiedlichen Schichten
- Faster R-CNN (Inception und ResNet)
 - ist ein Two-stage Detector: 1. ROIs mithilfe RPN oder SelSearch finden, darauf dann Classifier anwenden
- FRCNN

(In eval Ergebniss dann etwa so: SSD zu schlechte Performance und für Appl keine Realtime notwendig, FRCNN zu langsam, Faster R-CNN gute Mitte)
Hier ersten Durchlauf (mit Overfitting) darstellen

4.2.2 Regularisierung

Um das Overfitting zu vermeiden gibt es wie in 2.1.1 beschrieben verschiedene Möglichkeiten.
Untersucht wurde hier

- Augmentierung
- Early Stopping
- ... weitere z.B. L_1 , L_2

4.2.3 Training grayscale

Da die Kamera im Infrarot Modus ein Graustufen Bild mit nur einem Farbchannel liefert, muss dies für die Inferenz berücksichtigt werden.

Es ergeben sich hier mehrere möglichkeiten:

1. Normales (r, g, b) Netz
2. Ein Farbchannel (gr) Netz
3. Drei Farbchannel (gr, gr, gr)

Für 1. und 3. Müssten die bilder der Kamera vor der Inferenz auf 3 Farbchannel ($3 \times \text{grau}$) erweitert werden.

Um das Netz auf einen Channel zu trainieren wurde im config file ...

Um $3 \times \text{gray}$ zu trainieren wurden die Bilder in OpenCV in grau convertiert und wieder als jpgs abgespeichert.

Die Ergebnisse sind in Kapitel 5 dargestellt.

4.3 Parameter Optimierung

einstellungen im Config File

tensorflow graph oder plot zeigen

loss erklären (mit formel und für train und eval)

Kapitel 5

Evaluierung

5.1 Evaluierungs Metriken

mAP

- IoU Intersection over Unit: überlappung pred box mit ground truth box fläche beider boxen zusammen
- Confusion matrix
 - TP: True Positive (richtig auf ja getippt) (TP wird bei IoU \geq threshh. (übl 0.5) festgelegt)
 - TN: True Negative (richtig auf nein getippt)
 - FP: False Positive (fälschlicher ja getippt)
 - FN: False Negative (fälschlicherweise auf nein getippt)
- Precision: $TP / (TP + FP)$: möglichst viele richtige aus allen finden
- Recall: $TP / (TP + FN)$: möglichst **nur** die richtigen (sensitivity)
- AP = Precision-Recall verhältniss, genauer fläche unter kurve der beiden
- mAP für alle klassen gemittelt

Loss

kombination aus binärem classification log loss (cross entropy) für boxen und L1 smooth loss.

5.2 Ergebnisse

5.2.1 Modell Vergleich

hier modelle auf genauigkeit und geschwindig keit vergleichen und für nächste abschnitte eins auswählen

5.2.2 Regularisierung

ergibt dann zB early stoping und aug sind gleich.

daher wird im nächsten abschnitt geproüft wie sich die Modelle für Daten einer anderer Distribution verhalten.

Regularisierung	mAP	Loss
keine	1	2
Early Stopping	1	2
Augmentierung	1	2
weight decay	1	2

Tabelle 5.1: Regularization

5.2.3 Dataset Distributions

Da sowohl trainings als auch eval daten aus dem web bezogen wurden, also der gleichen distribution sind, diese aber nicht unbedingt den realen bedingungen entsprechen, wurde ein weiteres Evaluierungs Set mit Eigenen Aufnahmen erstellt, welche, da in der Umgebung (Wiltierpark in Reutlingen) aufgenommen, der tatsächlichen daten wahrscheinlichkeit eher entspricht.

ergebnisse tabellarisch vergleichen: handy vs orig für die in 5.2.2 beschriebenen Regularisierungs techniken

wenn für Augmentierung besser, heist das Augmentierung ist besser als early stopping für robustheit gegenüber umgebung (zb belichtungen)

Die Regularisierungen Early Stopping und Augmentierung wurden nun noch einmal mit einem eigenen Datenset evaluiert und miteinander verglichen.

Regularisierung	mAP_{orig}	mAP_{handy}	$Loss_{orig}$	$Loss_{handy}$
Early Stopping (100k steps)	0.6715	0.4265	0.6742	0.267
Augmentierung (200k steps)	0.6914	0.4537	0.6738	0.2503

Tabelle 5.2: Regularization

Wie zu erwarten unterscheiden sich die Loss Werte für das Origianel Eval Set wegen Early stopping nicht sehr. Für die Handy Bilder ist der Augmentierungs Loss etwas besser, was auf eine gr robustheit gegenüber Daten aus anderer Distribution zurückzuführen ist.

Die mAP Werte sind sowohl für original als auch für handy datenset bei Augmentierung deutlich besser, da sich wie in 5.1 der mAP erst später seinem Endwert annähert als der Loss.

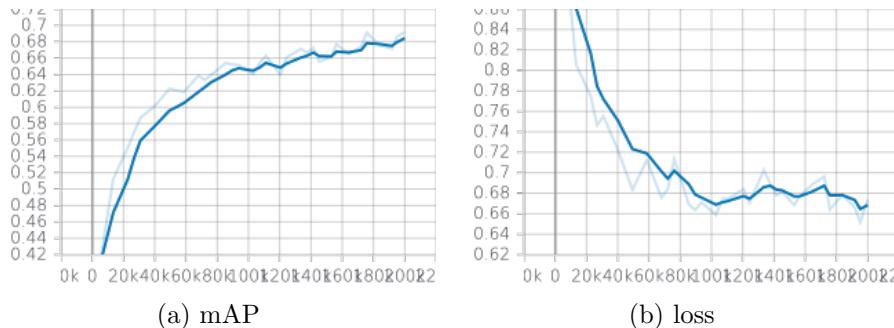


Abbildung 5.1: Loss und mAP für 200000 Steps

5.2.4 Graustufen/Infrarot Bilder

Da, wie die Ergebnisse in 5.2.2 und 5.2.3 gezeigt haben eine Augmentierung (welche Augmentierungen) die erfolgreichste Regularisierungs technik war, wurde für die Graustufen Bilder nur Auf Augmentierte Bilder mit faster Inception trainiert:

hier für die drei in 4.2.3 verwendeten modelle jwls für die in 5.2.3 beschriebenen eval daten sätze vergleichen.

Modell	Dataset	mAP	Loss
rgb	original	0.6556	0.1451
	handy	0.4155	0.2389
gray 1 channel	original	0.5625	0.1716
	handy	0.3226	0.2747
gray 3 channel	original	0.664	0.1653
	handy	0.438	0.2492

Tabelle 5.3: Grayscale

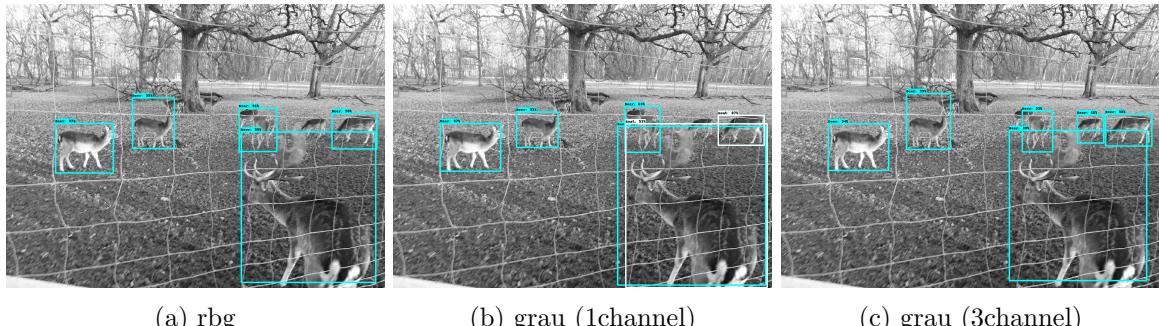


Abbildung 5.2: Vergleich der Inferenz von grau bildern für verschiedene Modelle

Diskussion des Ergebnisses: welchen einfluss haben Form und Farbe bei training, unnötig gelernte features für gray bilder? schnelligkeit?

Da es sich bei den convertierten graustufen bilder ja nicht um echte Graustufen bilder handelt, wurden Test Bilder von Alltags gegenständen mit der Für die Applikation verwendeten RaspiCam im Infrarot Modus aufgenommen. Um diese

(hat der Infrarot Modus mit Wärme, also lebendigkeit des Objektes zu zun??)

Um diese Zu testen wurde ein weiteres Netz auf diese Gegenstände (Gesicht, Hand,) trainiert und im folgenden auf den Datensatz true-ir evaluiert.

Kapitel 6

Entwicklung der Anwendung

In diesem Kapitel wird die Entwicklung der Anwendung als Autonomes Edge System auf dem Raspberry Pi zusammen mit dem Neural Compute Stick 2 und einer geeigneten Kamera beschrieben. Ebenso wird die Integration des trainierten Tensorflow Models in die Applikation sowie die Implementierung der Netzwerk Verbindung zu dem System beschrieben.

6.1 Aufbau

Die Anwendung soll auf dem ein Platinen Computer Raspberry Pi 4 Abbildung ?? laufen, an den die nötigen Komponenten angeschlossen werden. Dazu gehören der Neural Compute Stick 2, zur Ausführung der Inferenz, ein Kamera Modul, mit welchem die Bilder aufgenommen werden, sowie ein WiFi Stick und Powrebank.

Der Neural Compute Stick wird über USB angeschlossen und kann nach installation des OpenVino Toolkits 6.2 verwendet werden.

Bei der Kamere handelt es sich um ein Infrarot Fähiges *RaspberryPi Camera Module* welches zusammen mit zwei Infrarot LEDs montiert wird. 6.3



Abbildung 6.1: Raspberry Pi 4

- Netzwerkverbindzuung:
 - GSM Module
 - WiFi-Stick
- Powrebank/Akku aus Sp. Verbrauch von:
 - NCS2
 - Kamera
 - LEDs
 - WiFi Stick

6.2 OpenVino Toolkit

Die Implementierung der Inferenz des trainierten Models wurde mithilfe des OpenVino Toolkits vorgenommen, eine Anwendung zur Optimierung und Ausführung von CNNs auf Intel Hardware. Es vereinfacht und Optimiert damit die Verbindung zwischen Training des Models und bereitstellen in einer Anwender Applikation, wie in Abbildung 6.2 schematisch dargestellt.

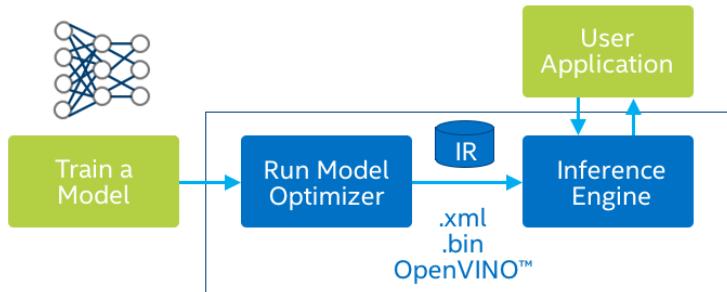


Abbildung 6.2: Workflow: OpenVino Toolkit

Das Toolkit besteht im Wesentlichen aus den zwei Komponenten *Model Optimizer* und *Inference Engine*

Mit dem Model Optimizer können Netze die in den Frameworks TensorFlow, Caffe, MXNet, Kaldi oder ONNX trainiert wurden in die von OpenVino verwendete Intermediate Representation des Modells gebracht werden.

Diese ist ein Framework unspezifisch Dateiformat, welches aus einer .xml Datei für die Struktur/Architektur des Modells und einer .bin Datei für die trainierten gewichten besteht.

Die InferenceEngine ist eine Runtime welche eine API für die Sprachen C++ und Python zur Integration und Nutzung der Inferenz in der Anwendung bereitstellt.

Dafür werden die IR Dateien des Models in ein Hardware spezifisches Plugin geladen. Dieses kann die User Applikation für die Inferenz von Image Classification, ObjectDetection sowie Instance Segmentation Modellen nutzen.

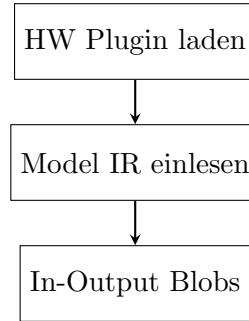
Implementierung

Die Implementierung der Inferenz wurde in Python vorgenommen.

Dafür waren folgende Schritte nötig:

1. HW Plugin laden
2. Model IR einlesen
3. In-Output Blobs allokieren
4. ausführbares Model laden
5. inferenz request abgeben
6. Bild als Array in Input Blob laden
7. Inferenz
8. Output verarbeiten, wieder zu Schritt 6

Blobs sind In-Output Tensoren



```

plugin = IEPlugin(device='MYRIAD')
net = IENetwork(model=model_xml, weights=model_bin)
input_blob = next(iter(net.inputs))
exec_net = plugin.load_network(network=net)
infer_request = exec_net.requests[request_id]
# bild mit opencv als numpy arra laden und von hwc nach nchw umstellen
res = exec_net.infer(inputs={input_blob: image})
# res enthaelt liste mit allen erkannten klassen auf dem Bild
# fuer Objekt Detection zusaetlich noch Bounding Box koordinaten
  
```

Die Inferenz kann entweder Synchron oder Asynchron ausgeführt werden. Der programmatische Ablauf der hier verwendeten asynchronen Inferenz ist im Folgenden als Pseudocode dargestellt.

```

while true do
    capture frame;
    populate Next InferRequest;
    start Next InferRequest; // asynchroner aufruf
    if wait for Current done then
        // wird in eigenem verarbeitet
        display Current;
    end
    swap Current and Next InferRequests;
end
  
```

Algorithm 1: Asynchrone Inferenz

Das Ergebnis eines InferRequest für Object Detection Modelle enthält eine Liste mit allen möglichen erkannten Objekten, jedes davon bestehend aus einem Array mit den Indices:

0. batch index
1. class label
2. Wahrscheinlichkeit
3. x_{min} Box Koordinate
4. y_{min} Box Koordinate
5. x_{max} Box Koordinate
6. y_{max} Box Koordinate

Mit über die Wahrscheinlichkeit ließen sich die Ergebnisse nach einem bestimmte Threshhold ausfiltern.

Die Box Koordinaten wurden in Prozent der Bild- Breit/Höhe angegeben wodurch sie wieder in die Original bild Größe für die Bounding Boxes übertragen werden können.

6.3 Raspberry Pi Kamera

Bei der Kamera handelt es sich um das OV5647 5MP Modul mit regelbarem Infrarotfilter. Zusammen mit zwei Infrarot LEDs von der Firma Quimat Abbildung ??

Wird der Infrarotfilter ausgeschaltet ist es durch die Infrarot LEDs mit 850nm welligen Licht möglich auch bei Dunkelheit Aufnahmen zu machen, die in Graustufen Werten dargestellt werden.

6.4 Server-Client-Connection

6.5 Anwendung gesamt

Da die Inferenz sehr rechenaufwendig ist, sollen die Frames der Kamera nur dann inferiert werden, wenn eine Bewegung stattfindet. Dafür wurde der Inferenz ein Bewegungsmelder vorgeschaltet. Dieser wurde mithilfe der Library OpenCV implementiert, indem zu Beginn des Kamereastrams ein Referenzbild gespeichert wurde, mit dem die aktuellen Frames verglichen werden. Ist der absolute Abstand der einzelnen Array Elemente/Werte der Bilder größer als ein bestimmter Threshold, wird dies als Bewegung gewertet.

Kapitel 7

Test und Validierung

Kapitel 8

Zusammenfassung und Ausblick

Die Zusammenfassung bildet mit der Einleitung den Rahmen der Arbeit. Sie greift zu Beginn die Aufgabenstellung auf und beschreibt dann die wesentlichen Punkte des Lösungsweges und die erzielten Ergebnisse kurz und knapp, so dass diese in kürzester Zeit erfasst werden können.

Anschließend werden noch kurz offene Punkte, Verbesserungen oder Weiterentwicklungen diskutiert.

Insgesamt sollten Zusammenfassung und Ausblick anderthalb Seiten nicht überschreiten. In der Regel ist eine Seite ausreichend.

Literaturverzeichnis

Anhang A

Beispiel für ein Kapitel im Anhang

A.1 Bsp für ein Abschnitt im Anhang