HAW Hamburg

Fußgängererkennung im Straßenverkehr mittels Deep Learning

Patrick Nagorski
patrick.nagorski@haw-hamburg.de
Department Informatik
Hochschule für Angewandte Wissenschaften Hamburg
Berliner Tor 7
20099 Hamburg

31. August 2018

Inhaltsverzeichnis

1 Einleitung

1.1 Problemstellung

In Städten mit größerem Verkehrsaufkommen ist die ständige Aufmerksamkeit von allen Verkehrsteilnehmern von hoher Wichtigkeit. Sind Autofahrer unaufmerksam und beispielsweise durch Smartphones, Essen oder Schminken abgelenkt, ist die Wahrscheinlichkeit eines Verkehrsunfalls höher. So können Fußgänger übersehen und verletzt werden.[1]
Durch neue Technologien können solche Unfälle vermieden werden. Das Erkennen von Hindernissen durch Sensoren und ein anschließendes automatisches Abbremsen des Fahrzeugs ist in neueren PKWs bereits der Standard. Diese Techniken können dazu genutzt werden zuverlässige Handlungsprognosen bei Gefahren aufzustellen, wie beispielsweise das Analysieren von herannahenden Fußgängern durch die zeitliche Veränderung des Standortes [2]. Außerdem muss berücksichtigt werden, ob sich ein Fußgänger nur auf dem Fußgängerweg befindet oder die Straße überquert [3]. Hektische und plötzliche Bewegungen und eine große Anzahl von Fußgängern machen dies komplizierter.
LKWs besitzen bekanntermaßen einen sehr großen toten Winkel, wodurch andere Verkehrsteilnehmer, wie Fußgänger, Fahrradfahrer oder Motorradfahrer nicht gesehen werden und so ein Unfall entstehen kann [4]. Hier kann durch dieselben Technologien ein System entwickelt werden, welches eine Unterscheidung treffen kann, um was für einen Verkehrsteilnehmer es sich handelt und so sichere Handlungsprognosen trifft. Solche Systeme könnten in vielen weiteren Bereichen ebenfalls von Nutzen sein.

1.2 Motivation

Ziel ist die Entwicklung eines Systems, dass auf das Erkennen eines Objektes reagiert. Beispielsweise das Erkennen von Fußgängern, die sich auf der Straße oder am Straßenrand befinden. Die Fußgängererkennung ist eine Unterkategorie der Objekterkennung, in der es das Ziel ist ein Bild bzw. Video zu nehmen und die Präsenz von Fußgängern in dem Bild zu erkennen [2]. Es soll eine Differenzierung von Voraussagen getroffen werden und so vom System entschieden werden, ob Gefahr besteht oder nicht. Diese Methode soll für PKWs, LKWs und Züge geeignet sein. Zudem wäre es möglich die Methode auf das Erkennen von Tieren zu spezialisieren. So könnten Wildunfälle ver-

mieden werden, aber auch Bestände gefährdeter Tierarten kontrolliert werden. Da vor allem nachts die Sicht bei menschlichen Augen eingeschränkt ist, sind Kameras und Sensoren notwendig, um Gefahren rechtzeitig zu erkennen. Die Technologie soll auf der Anwendung von künstlichen neuronalen Netzen mittels Deep Learning Methoden basieren, da diese in der Vergangenheit die besten Ergebnisse geliefert haben [5].

2 Grundlagen und Methodik

2.1 Technologien

Für die Entwicklung einer entsprechenden Methode zur Erkennung von Fußgängern im Straßenverkehr wurden die beiden wichtigen Technologien CNN (Convolutional Neural Network) und LSTM (Long short-term memory) verwendet.
Bei CNN handelt es sich um Faltungsnetzwerke. Diese werden aufgrund der hohen Performance und der Flexibilität hauptsächlich für für Objekterkennung verwendet [5]. Ein weiterer Vorteil ist, dass räumliche Zusammenhänge der Bild-Dateien erhalten bleiben. Sie sind robust gegen Rotationen, Translationen sowie Skalierungen. Durch Data Augmentation können zusätzliche Trainingsdaten für CNNs erzeugt werden. CNNs werden üblicherweise überwacht trainiert. Während des Trainings wird dabei für jeden gezeigten Input der zugehörige One-Hot-Vektor bereitgestellt. Mittels Backpropagation wird der Gradient jedes Neurons berechnet und die Gewichte in Richtung des steilsten Abfalls der Fehleroberfläche angepasst.[6]
Bei LSTM handelt es sich um ein Rekurrentes Neuronales Netz (RNN), welches für zeitlich abhängige Daten angewendet wird. Es ermöglicht sequenzielles Lernen und langfristige Abhängigkeiten sind innerhalb einer Sequenz behandelbar. Im Gegensatz zu regulären RNNs hat LSTM die Vorteile, dass sowohl kurze, als auch lange Zeitabhängigkeiten verarbeitet werden können.[6]
Diese beiden Technologien können als Hybrid-Netz angewendet werden, wodurch sowohl die räumlichen Zusammenhänge der CNN und die zeitliche Abhängigkeit von LSTM Netzen betrachtet werden können [7]. Ein beispielhaftes Hybridnetz ist in Abbildung 1 zu sehen.
Die Problemstellung lässt sich der Many-to-One Kategorie zuordnen, da immer erst nach einer bestimmten Anzahl von Zeitschritten eine Klassifikation stattfinden soll [8].

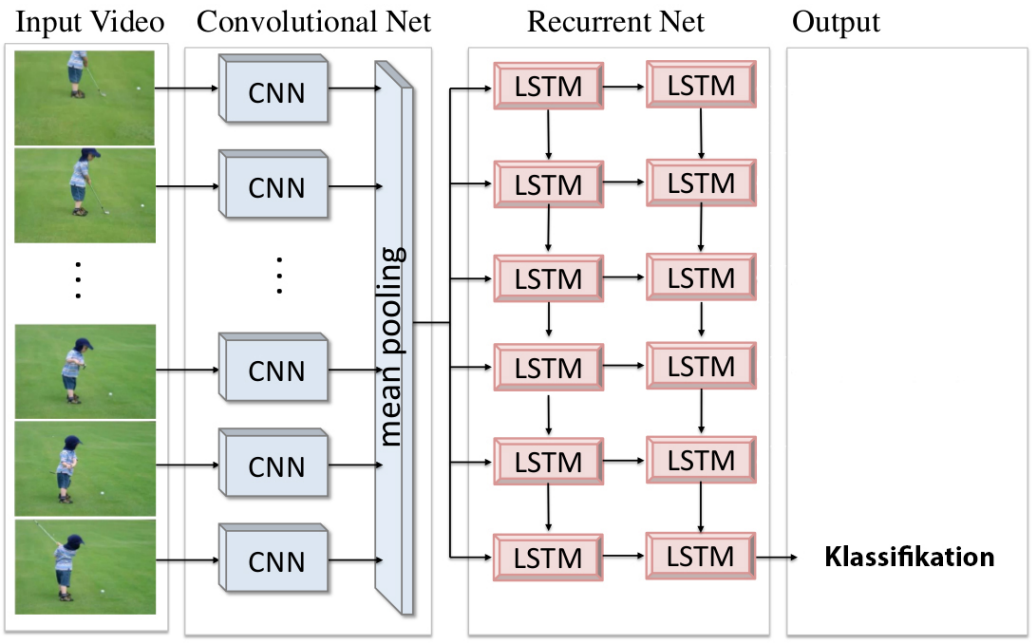


Abbildung 1: Aufbau eines hybriden CNN / LSTM-Netzes. In Anlehnung an [9]

2.2 Datasets

Für das Training von neuronalen Netzen wurden für diese Arbeit die folgenden zwei Datasets betrachtet.

Caltech Pedestrian Dataset:
Das Caltech Pedestrian Dataset ist ein sehr bekanntes Dataset mit Fußgängern. Es beinhaltet 10 Stunden Videomaterial (640x480, 30 Hz). Dieses Videomaterial wurde aus einem fahrenden Auto aufgenommen und enthält 250.000 Einzelbilder. Im Dataset sind 2300 einzigartige Fußgänger dargestellt. Ein Ausschnitt des Datasets ist in der Abbildung 2 zu sehen. In der Abbildung befinden sich Fußgänger auf der Straße, die mit einem grünen Rechteck als Fußgänger markiert sind.[10]

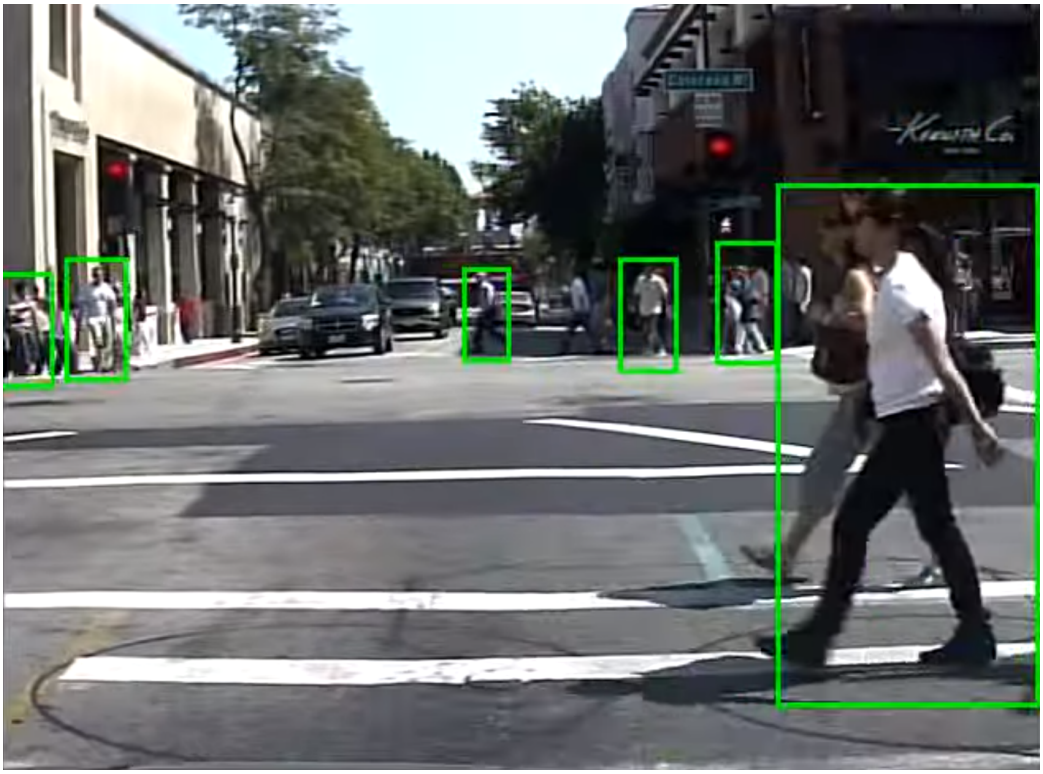


Abbildung 2: Ausschnitt aus dem Caltech Pedestrian Dataset.[10]

BDD100K Dataset:
Das BDD100K Dataset ist das größte vorhandene Dataset, welches ebenfalls aus einem Auto aufgenommene Videosequenzen enthält. Es ist seit 2018 zugänglich und umfasst 100.000 HD Videos mit einer Gesamtlänge von 1.100 Stunden. Verfügbar sind unterschiedliche Tageszeiten, Wetterbedingungen und Fahrszenarios. Zusätzlich ist eine Straßenobjekterkennung vorhanden, die 100.000 Bilder mit 2D-Rahmen umfasst. So sind Ampeln, Straßenschilder, Personen, Fahrräder, Autos und viele weitere Objekte differenziert. In Abbildung 3 sind verschiedene Ausschnitte zu unterschiedlichen Tageszeiten und Wetterbedingungen dargestellt.[11]

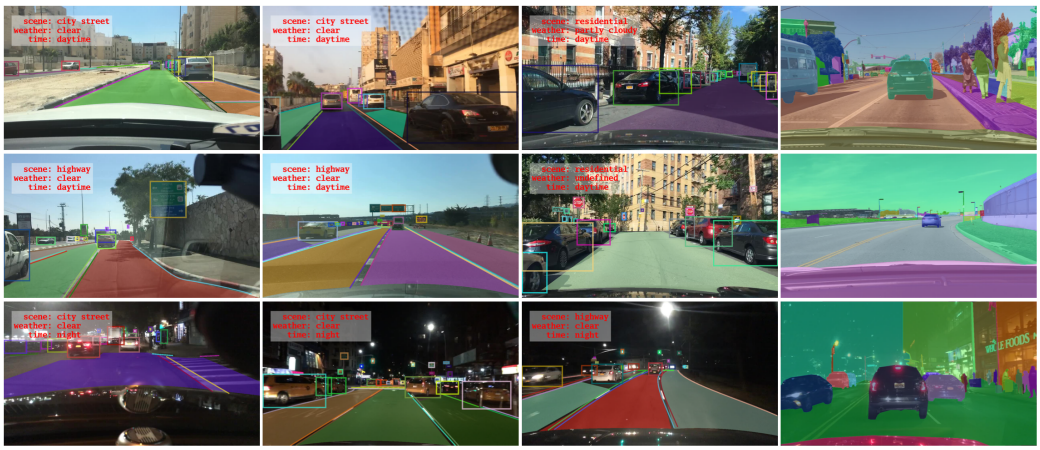


Abbildung 3: Ausschnitte aus dem BDD100K Dataset.[11]

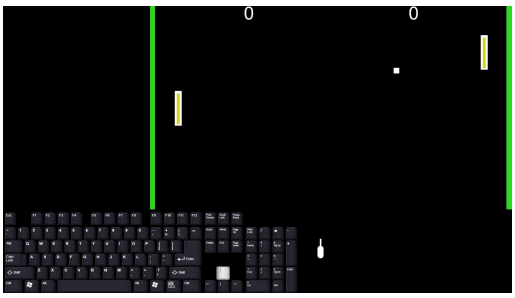
2.3 Konferenzen und Journale

Fußgängererkennung spielt sowohl bei Fahrassistenzsystemen, als auch beim Gebiet des autonomen Fahrens eine bedeutende Rolle. Aus diesem Grund gibt es viele Konferenzen, die auf dieses Themengebiet ausgelegt sind.
Die TU-Automotive Detroit ist die weltweit größte Konferenz und Ausstellung, die die Zukunft der vernetzten und autonomen Fahrzeuge zeigt. Zusätzlich werden Fahrerassistenzsysteme vorgestellt.[12]
Das Autonomous Vehicle Software Symposium diskutiert die Herausforderungen bei der Programmierung autonomer Fahrzeugsoftware. Es wird besprochen, wie die Entwicklungszeiten reduziert und gleichzeitig die Sicherheiten erhöht werden können. Außerdem werden KI-Herausforderungen und Entscheidungsprozesse thematisiert.[12]
Das Automated Vehicles Symposium versammelt Industrie, Regierung und Wissenschaft aus der ganzen Welt, um komplexe Technologie, Operationen und politische Probleme bezüglich des autonomen Fahrens anzugehen. Ziel ist es darüber zu informieren, den Fortschritt in Richtung Sicherheit und automatisierte Mobilität voranzutreiben.[12]
Das SAE 2018 ADAS to Automated Driving Symposium unterstützt die Automobilbranche bei der Einführung von Advanced Driver Assist Systems (ADAS) und vollautomatischem Fahren. Der Fokus liegt auf automatisiertem Fahren und aktive Sicherheit.[12]
Das International Journal of Vehicle Autonomous Systems (IJVAS) ist eine

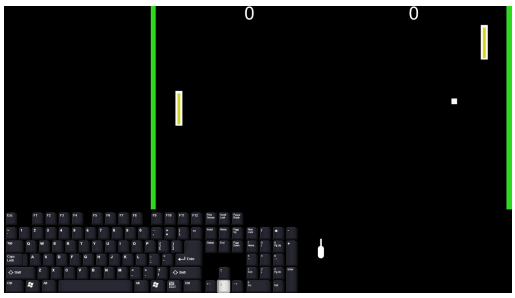
etablierte internationale Referenz auf dem Gebiet der Forschung und Entwicklung von autonomen Systemen für Fahrzeuge.
Solche Systeme zielen darauf ab, dass Unfälle vermieden werden, dass das Reiseerlebnis verbessert wird, indem die Insassen von Fahr- / Navigationsarbeiten entlastet werden, die Gesamtfahrzeuganzahl reduziert wird und einige der mit dem Autofahren verbundenen Dienste / Infrastruktur beseitigt werden.

2.4 Pong

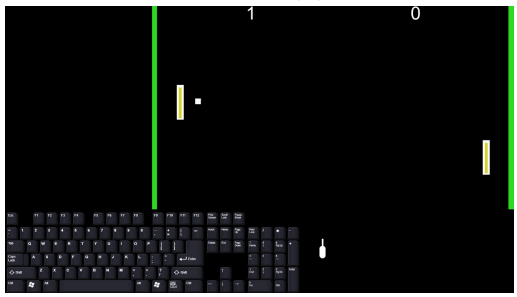
Die Machbarkeit der Anwendung der genannten Technologien für eine Fußgängererkennung wurde anstelle mit den Datasets zunächst mittels Pong [13] getestet. Die Anwendung von Pong als Einstiegsprojekt ist weniger komplex, als die Analyse von Verkehrssequenzen und es kann ebenfalls der Many-to-One Kategorie zugeordnet werden, da nach n Zeitschritten eine Aktion stattfinden soll bzw. die Situation des Balls zugeordnet werden soll. Würden nur die Einzelbilder anstatt eine Sequenz von Bildern betrachtet werden, wäre nicht sichtbar, ob sich der Ball zum Schläger bewegt oder sich vom Schläger entfernt. Dies wäre bei der Fußgängererkennung ebenfalls relevant, da es unterschieden werden sollte, ob der Fußgänger den Fußgängerübergang erst betritt oder bereits verlässt.
Die Datenerzeugung erfolgte selbst, sodass keine Abhängigkeit von Daten aus dem Netz vorhanden war. Die Daten wurden durch das Spielen eines menschlichen Spielers gegen einen Computergegner generiert.
Zu Beginn wurden die Daten vorverarbeitet. Hierfür wurde das 8:35 min lange Video in Einzelbilder aufgeteilt. Insgesamt entstanden so 30.939 Einzelbilder. Diese wurden anschließend als Graubilder eingelesen und weiterverarbeitet, da dies die Anzahl der Informationen verringert (im Vergleich zu RGB Bildern) und das Labeln vereinfachte.
Anschließend wurde auf zwei unterschiedliche Weisen gelabelt. Die erste Art berücksichtigte die Aktion, welche vom menschlichen Spieler durchgeführt wurde (UP, DOWN, NONE). Dabei wurde während des Spiels eine Tastatur auf dem Bildschirm eingeblendet, welche darstellte, welche Tasten gedrückt wurden. Dies wurde anhand des Grauwertes an den Positionen der Pfeiltasten der Tastatur ermittelt. Gedrückte Tasten haben einen hohen Grauwert, nicht gedrückte Tasten dagegen einen niedrigen Grauwert. Drei unterschiedlich klassifizierte Einzelbilder sind in Abbildung 4 zu sehen. Dies ist an den hellen gedrückten Pfeiltasten ersichtlich.



(a) Pong Beispiel mit Label UP.



(b) Pong Beispiel mit Label DOWN.



(c) Pong Beispiel mit Label NONE.

Abbildung 4: Pong Beispiel mit den Labels (UP, DOWN, NONE).

Die zweite Label-Art berücksichtigt die Situation des Balls auf dem Spielfeld (HIT, MISS, NEUTRAL). Hierfür wurden drei Situationen betrachtet: Der Ball wurde vom Schläger getroffen, der Schläger verfehlt den Ball und der Ball bewegt sich allgemein auf dem Spielfeld, wobei er sich nicht in der Nähe des Schlägers des menschlichen Spielers befindet. Dies wurde in Abbildung 5 dargestellt.
Um den Zusammenhang zu verdeutlichen wurden in dem Einzelbild der Abbildung 5 gestrichelte Linien hinzugefügt. Die Gelb gestrichelte Linie entspricht den möglichen Positionen unmittelbar vor dem Schläger des menschlichen Spielers, die Weiß gestrichelte Linie entspricht den möglichen Positionen des Schlägers des menschlichen Spielers und die Magenta gestrichelten Linien entsprechen den möglichen Positionen hinter dem Schläger des menschlichen Spielers.

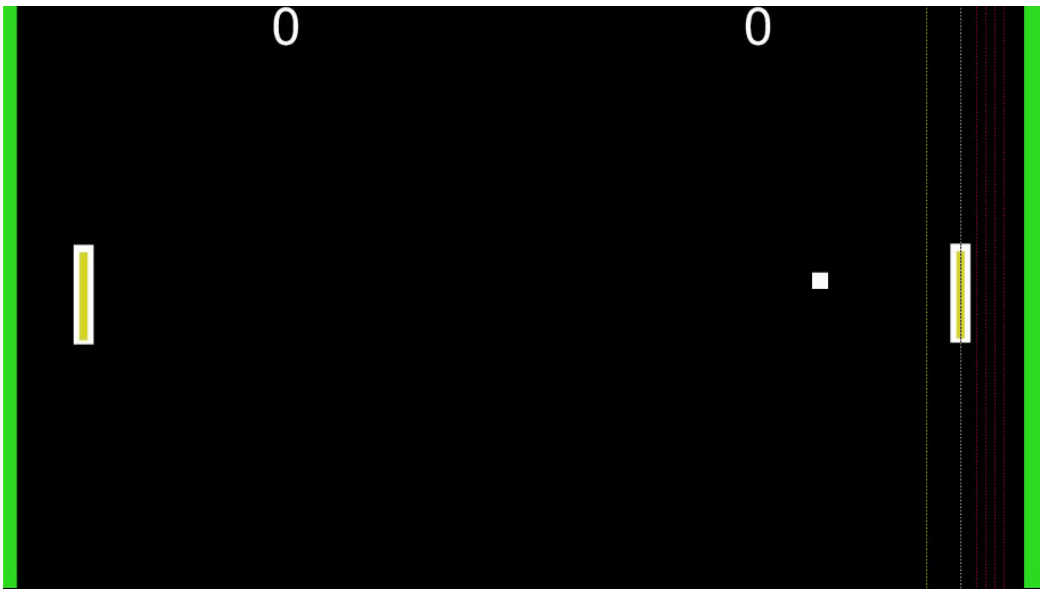


Abbildung 5: Pong Beispiel mit den Labels (HIT, MISS, NEUTRAL).

Die Grauwerte der Positionen der gestrichelten Linien wurden geprüft und verglichen. Da das Spielfeld einen niedrigen Grauwert hat und der Ball einen hohen Grauwert hat, kann dadurch überprüft werden, ob sich der Ball in den Positionen der gestrichelten Linien befindet.
Es wurden folgende Szenarien festgestellt und verglichen:

MISS, wenn:

- der Grauwert an einer Position einer Magenta gestrichelten Linie größer 0 ist
- der Grauwert an der n-ten Position der Gelb gestrichelten Linie größer 0 ist und der Grauwert an der n-ten Position der Weiß gestrichelten Linie kleiner 1 ist

HIT, wenn:

- der Grauwert an der n-ten Position der Gelb gestrichelten Linie größer 0 ist und der Grauwert an der n-ten Position der Weiß gestrichelten Linie größer 0 ist

NEUTRAL, wenn:

- Grauwerte auf allen Positionen der Gelb gestrichelten Linie, der Weiß gestrichelten Linie und der Magenta gestrichelten Linien sind 0

Das n entspricht den Werten von 0 bis 508 (Höhe des Spielfeldes).

Nach dem Labeln wurden die Einzelbilder zugeschnitten, damit nur das Spielfeld vorhanden war. Die Tastatur wurde nur für das Labeln (von Art 1) benötigt.
Da die Auflösung der Einzelbilder mit 910 x 510 noch recht groß ist, wurde diese auf 91 x 51 reduziert. Mit dieser Auflösung waren alle Elemente des Spielfeldes immer noch sichtbar und die Einzelbilder bestanden aus deutlich weniger Pixeln, sodass es die Trainingszeit des neuronalen Netzes deutlich verringert werden konnte.
Das für diese Arbeit verwendete LSTM-Netz benötigt Vektoren als Input. Daher wurden die Arrays dementsprechend angepasst. Zusätzlich wurden die Einzelbilder in Pakete aufgeteilt, die je vier Einzelbilder enthielten, wobei das letzte Label nach vier Einzelbildern galt. So fand, wie in Abbildung 6 zu sehen, jeweils nach vier Einzelbildern eine Kategorisierung statt.
Die Pakete wurden zusätzlich immer um einen Schritt verschoben, um das Muster 1, 2, 3, 4; 2, 3, 4, 5; ... zu erhalten. Durch die Überschneidungen der Pakete konnten mehr Trainingsdaten erhalten werden, als ohne eine Überschneidung.

```
(996, 4, 1400)
[[[ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]]

 [[ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]]

 [[ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]]

 ...

 [[ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   0.   1. 109.]]

 [[ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   0.   1. 109.]]

 [[ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   0.   1. 109.]
  [ 95.   0.   1. ...   1.   0. 109.]]]
```

(a) Vierer Pakete der Einzelbilder.

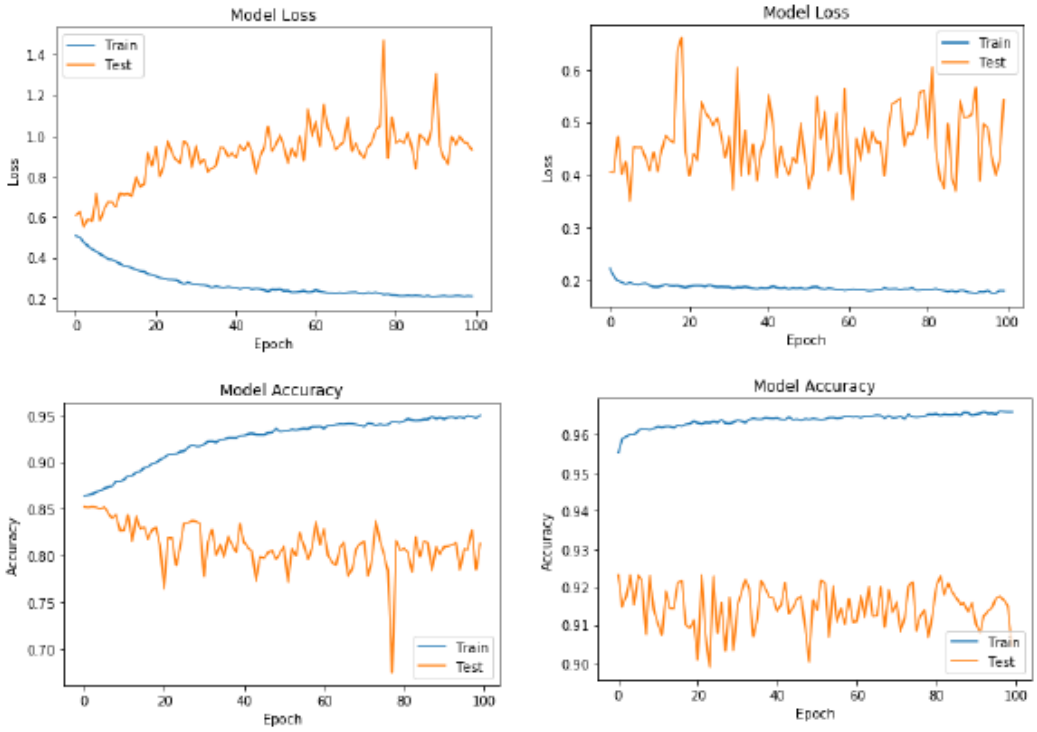
```
(996, 3)
[[1. 0. 0.]
 [1. 0. 0.]
 [1. 0. 0.]
 ...
 [1. 0. 0.]
 [0. 0. 1.]
 [1. 0. 0.]]
```

(b) Zu den vierer Paketen der Einzelbilder zugehöriges Label.

Abbildung 6: Loss und Accuracy Graphen der beiden Label-Arten.

3 Bisherige Ergebnisse

Für das Training wurden 19996 Trainingsbilder und und für das Testen des trainierten Netzes wurden 9996 Testbilder verwendet. Um die Genauigkeit der Klassifizierung des trainierten Netzen auszuwerten, wurden mit Hilfe des Programms Grafiken erstellt (Abbildung 7). Diese zeigen durch den Accuracy-Wert die Genauigkeit und durch den Loss-Wert an, wie sich die Genauigkeit im Vergleich zur vorherigen Epoche entwickelt hat.
Die $Label_1$ Trainings Accuracy geht kontinuierlich hoch und nähert sich 95% an. Die Test Accuracy hat Schwankungen im Verlauf der Epochen, befindet sich aber gegen Ende zwischen 80% und 83%. Der Verlauf wird ebenfalls auf der Loss Kurve deutlich, während bei den Trainingsdaten die Kurve sinkt, bis sie ungefähr 0.2 erreicht, springt sie bei den Testdaten hin und her.
$Label_2$ Trainings und Test Accuracy Verlauf ähnelt dem von $Label_1$, nur dass die Training Accuracy die 96% überschreitet. Die Test Accuracy pendelt sich zwischen 91% und 92% ein.
Die niedrigere Accuracy von $Label_1$ im Vergleich zu $Label_2$ lässt sich erklären, da das Labeln vom Spielstil eines menschlichen Spielers handelt und dieser Abweichungen im Spielverlauf hat. Dies hat einen Einfluss auf die Accuracy, da das System kleine Abweichungen bereits als Fehler interpretiert, obwohl diese keinen negativen Einfluss aufs Spielgeschehen haben müssen.



(a) Art 1 (DOWN, UP, NONE).

(b) Art 2 (HIT, MISS, NEUTRAL).

Abbildung 7: Loss und Accuracy Graphen der beiden Label-Arten.

4 Ausblick

Die bisher erzielten Ergebnisse weisen Genauigkeiten zwischen 80% und 90 % auf. Um die Genauigkeiten zu erhöhen werden bessere Trainingsdaten benötigt. Außerdem kann die Genauigkeit des Netzes durch das Anwenden eines CNN-LSTM-Hybrid-Netzes erhöht werden. Dies soll der nächste Schritt des Projektes beinhalten.
Daraufhin sollen an Stelle der Pong-Datasets, die hier vorgestellten Datasets verwendet werden. Anschließend soll eine Konkretisierung des Themas im Bezug auf das Thema der künftigen Masterarbeit folgen. Da es für PKW bereits viele Assistenzsysteme zur Fußgänger- und Verkehrszeichenerkennung gibt, soll bei der Auswahl des Themas für die Masterarbeit auf andere Anwendungen dieser Technik bezogen werden.
Beispielsweise besitzen Nachrüstsysteme zur Fußgängererkennung für LKWs eine Fehlerrate von 10% [14]. Lediglich ein auf Radartechniken basiertes System, welches bei der Produktion bereits verbaut sein muss, weist geringere Fehlerraten auf [14]. So könnte eine Technik entwickelt werden, die auf der Aufnahme von Bild- und Videosequenzen und der anschließenden Objekterkennung, sowie Warnung des Fahrers beruht.

Literatur

[1] DEKRA zu Ablenkung am Steuer durch Smartphones: https://www.dekra.de/de/dekra-zu-ablenkung-am-steuer-durch-smartphones/, 2017, Zugriffsdatum: 03.08.2018

[2] A. Variyar, Application of Convoluted Neural Networks for Pedestrian Detection, 2016

[3] E. Rehde et al., Pedestrian Prediction by Planning using Deep Neural Networks, 2017

[4] S. Summerskill, R. Marshall, Understanding direct and indirect driver vision from heavy goods vehicles, 2016

[5] X. Zhao et al., A Faster RCNN-based Pedestrian Detection System, 2016

[6] A. Meisel, Vorlesungsfolien - Modellierung dynamischer Systeme, 2018

[7] J. Brownlee, CNN Long Short-Term Memory Networks, 2017: https://machinelearningmastery.com/cnn-long-short-term-memory-networks/, Zugriffsdatum: 03.08.2018

[8] The Unreasonable Effectiveness of Recurrent Neural Networks, 2015: http://karpathy.github.io/2015/05/21/rnn-effectiveness/, 03.08.2018

[9] S. Venugopalan et al., Translating Videos to Natural Language Using Deep Recurrent Neural Networks, 2015

[10] P. Dollar et al, Pedestrian Detection: An Evaluation of the State of the Art, 2012

[11] F. Yu et al., BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling, 2018

[12] Top Autonomous Vehicle Conferences to attend in 2018–19, https://hackernoon.com/top-autonomous-vehicle-conferences-to-attend-in-2018-19-d3a526a41a9a, Zugriffsdatum: 03.08.2018

[13] Pong2: https://pong-2.com/

[14] D. H. Freedman, Self-Driving Trucks Tractor-trailers without a human at the wheel will soon barrel onto highways near you. What will this mean for the nation's 1.7 million truck drivers?, MIT Technology Review 03-04/2017, 2017

[15] F. Chollet, Deep Learning with Python, Manning, 2017

[16] S. Yin et al., Multi-CNN and Decision Tree Based Driving Behavior Evaluation, 2017