**Proposal der Masterarbeit:**

1. Thema Deiner Masterarbeit
2. Hintergrund
3. Forschungsthema
4. Aktueller Stand der Forschung
5. Thesen und Ziele
6. Methodologie
7. Planung der Arbeit

**1. Motivation des Themas**

Computer Vision bezeichnet eine Technik der künstlichen Intelligenz, die es ermöglicht, Bilder zu analysieren, die von einem optischen Aufnahmegeräten wie einer Kamera aufgenommen wurden. Konkret handelt es sich bei Computer Vision (CV) um ein auf KI basierendes Werkzeug, das ein Bild erkennen, verstehen und die daraus resultierenden Informationen verarbeiten kann. Für viele ist Computer Vision das KI-Äquivalent zu den menschlichen Augen und der Fähigkeit unseres Gehirns, die wahrgenommenen Bilder zu verarbeiten und zu analysieren. Die Nachahmung des menschlichen Sehens durch Computer ist übrigens eines der großen Ziele der Computervision und erfordert das Aufbauen robuster NN-Modelle (wie DNN-Modelle) mit vielen Parameter. Oft werden DNN-Modelle für diese Aufgabe genutzt. Es gibt heute zahlreiche Anwendungsbereiche, wo Computer Vision zu Einsatz kommt. Darunter zählen zum Beispiel die Automobilbranche mit dem Aufkommen selbstfahrender Autos, die Bilder von der Straße erkennen können, sowie die Filmbranche mit Motion Capture, Gesichtserkennungssysteme in Smartphones usw.

Für viele ML-Systeme ist das Training nur die halbe Miete. Nachdem ein Modell trainiert wurde, muss es auch tatsächlich genutzt werden, um Vorhersagen und/oder Entscheidungen zu treffen. Dies wird in der Regel als Inferenz bezeichnet. Diese wird durch unterschiedliche Metriken evaluiert. Darunter zählen zum einen die accuracy des trainierten Modells auf eingehende Anfragen, die Latenzzeit (Wartezeit zwischen die Ausführung einer Anfrage und dem Vorhersagen), die Modelgröße sowie der Durchsatz, also wie viele Vorhersagen pro Sekunden aus dem System herausgeholt werden können, wenn es einen großen Stapel von Abfragen gibt, für die eine Vorhersage bestimmt werden soll. Auf der Hardwareebene sind ebenfalls Evalierungsmetriken wie Energien- und Leistungskonsum sowie Speicherbedarf und Deployment-Kosten auslandgenend für das allgemeine Deployment eines Model in ein gegebenes System.

**2. Punkt: Das Thema Deiner Masterarbeit und Hintergrund**

Das von Dir verwendete [Thema der Masterarbeit](https://gwriters.de/blog/masterarbeit-thema) ist ein vorläufiges. Doch wie so oft ist der Effekt der Gewohnheit groß und die Erfahrung zeigt, dass das zu Anfang provisorisch Ausgewählte bestand hat. Wichtig ist, dass er das sagt, worum es geht. Nicht poetischer Gehalt der Sprache ist gefragt, sondern die Sachlichkeit der Begriffe. Möglichst genau und konkret soll auf den Punkt gebracht werden, worum es in der Arbeit gehen wird.

**2. Punkt: Hintergrund (Problemstellung)**

[Wie bist Du auf die Fragestellung gekommen?](https://gwriters.de/blog/abschlussarbeit-thema) Warum interessiert es Dich so? Welche anderen Fachgebiete sind für das Thema relevant? Warum ist diese Frage wichtig und wieso hat sie noch niemand gestellt oder gut beantwortet? Und zum Schluss: Wer ist Dein Zielpublikum? Für wen ist die Beantwortung der Frage wichtig?

Ein besonderer Vorteil bei der Verwendung von DNN-Modellen ist vor allem auf die Skalierbarkeit zurückzuführen, mit der große Datenmengen kodiert und Milliarden von Modellparametern gesteuert werden können.[[1]](#footnote-1)

Die große Anzahl an Modellparametern macht Modelle rechenintensiv. Sie werden aus diesem Grund in Industrien, wo sie eingesetzt werden, oft auf extern gehosteten GPU-Servern deployed. Dabei ist die Übertragung der Bilderdaten auf externe Server jedoch rechtlich risikovoll und zeit- und kostenintensiv, da dazu hohe Netzwerkverkehr, Latenzzeiten und Bedenken zum Datenschutz getriggert werden. Der Einsatz von Computer Vision wird dadurch auf ressourcenarmen Edge-Geräten unpraktisch.

Aufgabenstellung in der Masterarbeit ist die Komprimierung und Bereitstellung von Deep Neural Networks Modellen, damit diese in industriellen Anwendungen schneller und effizienter ausgeführt werden können, dies auch auf Embedded Systemen zur Bildverarbeitung (NVIDIA Jetson Edge Devices).

**3. Punkt: Forschungsthema - Übersicht bestehender Ansätze und offene Fragen**

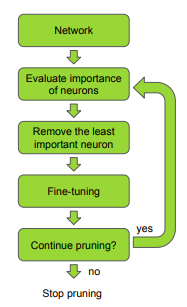
Hier sind wir nun im Kern Deines Masterarbeit-Proposals angekommen. Es geht darum zu beschreiben, was das Forschungsgebiet genau ist, was die Frage ist und wie sie verstanden werden soll. An dieser Stelle ist es auch wichtig, eventuell notwendige Begriffsabklärungen zu leisten und das Themengebiet in mehrere Fragen aufzuschlüsseln, damit die Komponenten, die für Dich das Thema ausmachen, einsichtig werden. Du kannst Dir das wie eine grobe [Gliederung](https://gwriters.de/blog/aufbau-wissenschaftliche-arbeit) vorstellen, in welcher die einzelnen Punkte dargelegt und erklärt werden.

Da Deployment sowie Kompression von DNN-Modellen wird in der Regel durch unterschiedliche Techniken erreicht. Dabei umfasst der Lösungsleitfaden Antworten auf Fragen zu gebrauchten Architektur, Konzepten & Tools und Implementierungsvorgehensweisen sowie zu Datenerfassung, -verarbeitung und möglichen Deployment-Environments. Zu Kompressionstechniken zählen unter anderem das Pruning und die Quatisation sowie die Interference Runtimes und die Knowledge Distillation.

Pruning Quantization:[[2]](#footnote-2)

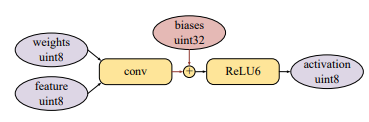
Das sind zwei Methoden zur optimierten Modellkompression. Die Optimierung handelt hier von der Anwendung minimierter Rechenressourcen und reduzierter Energiekosten, wobei weniger Verlust und mehr accuracy erstrebt werden. Wir vergleichen Pruning-Techniken und beschreiben Kriterien, die zur Entfernung redundanter Berechnungen verwendet werden.

* Pruning:



*Pruning wird als statisch kategorisiert, wenn es offline durchgeführt wird. Ein dynamisches Pruning wird zur Laufzeit durchgeführt. Die Unterscheidung beider Puning-techniken basiert auf Kriterien zur Entfernung redundanter Berechnung sowie auf Untersuchung einiger DNN-Komponenten wie z.B. Elemente, Filter, Layer, Netze, Kanäle, etc. (ToDo: Beschreibung)[[3]](#footnote-3)*

* Low-rank factorization
* Quantisation:

*Hier wird durch die Verringerung der Genauigkeit eines Datentyps die Reduzierung der Berechnungen erzielt. Man uunterscheidet zwischen uniformer und nicht uniformer Quantisation (ToDo: Beschreibung) [[4]](#footnote-4)*

Interference Runtimes:

*Beschreibung*

* + *NVIDIA TensorRT*
  + *Apache TVM*
  + *ONNX*

Knowledge Distillation:

*Hier lernt ein kleines Student-DNN-Modell von einem größeren Lehrer-DNN-Modell. Das Ganze beruht dabei auf einer Lehrer-student-Architektur. Die Destillation des Wissens erfolgt durch die Bildung von Wissenskategorien und Trainingsschemata. (ToDo: Beschreibung)*

Nur beide letztere Techniken werden im Rahmen der vorliegenden Arbeit erforscht.

Fragestellungen:

- Was sind die besonderen Eigenschaften von KD und wie funktioniert das?

- Auf welche Modelle werden die jeweiligen Kompressionsmethoden verwendet?

- Wie sehen diese algorithmisch aus und wie werden sie implementiert?

- Was ist das Problem mit bestehenden Lösungen dieser Art?

**5. Design des eigenen Einsatzes**

Auch hier ist es wichtig, dass Du Dich daran erinnerst, dass Du ein vorläufiges Statement verfasst. Es ist nicht in Stein gemeißelt, sondern kann sich im Laufe des Schreibens der Masterarbeit noch verändern. Aber für die Vermittlung ist eine ganzheitliche Sicht auf das Thema essenziell und darum auch Deine Vorstellung, wohin es Dich führen soll. Das also bedeutet: Was sind Deine Thesen? Wie bauen sie aufeinander auf? Welchen Nutzen ziehst Du aus ihnen? Wie oder woran kannst Du feststellen, dass Du Dein Ziel erreicht hast? Welche neuen Erkenntnisse werden dadurch erwartet? Und wer kann einen Nutzen daraus ziehen? Was sind die praktischen Anwendungsmöglichkeiten?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Meilenstein | Beschreibung | Dauer (Wochen) | Ziel | Vorgehen |
| 1 | Grundlage : Wissensphase bzw. Vorbereitungsphase | 4 | State of the Art herausfinden, Problem konfrontieren und Aufgabe erkennen und beschreiben | * Literaturrecherche * Verstehen und Beschreibung 3-5 Lösungssätze * Proposal * Auseinandersetzung mit gebrauchten Hardware (technische Beschreibung) |
| 2 | Grundlage 2: Konzeptphase | 4 | Implementierungskonzept erstellen: Design eines automatisierten Deployment Prozesses | * Struktur zur Implementierung der ausgewählten Lösungssätze verstehen * Algorithmische Lösung formulieren |
| Anmeldung | beim Prüfungsamt | (nach 8 Wochen, | am 18.04.2022) |  |
| 3 | Implementierung | 16 | Implementienrungspapeline (Zweiwöchige Terminabstimmung mit Betreuer) | Implementierung des designed Depoyment process. (Nachbauen Knowledge Distillation und Interference Rutimes) |
| 4 | Bugs fixing | 2 | Eventuelle Fehler im Code beheben | * Bugs fixing auf einzelnen Modulen testen * Bugs fixing auf Systemebene |
| 5 | Test und Evaluation | 4 | Hier werden Experimente geplant und durchgeführt. Es wird am Ende eine Evaluation der Experimente gemacht. Wichtig ist hier, dass der Accuracy bei der Komprimierung durch Einsatz der unterschiedlichen Methoden bestimmt wird, wobei ebenfalls die Laufzeit berücksichtigt wird. | * Experimente planen * Auswahl eines trainierten Modells (ReseNet8, MobileNet, …) mit bestimmter Aufgabe (Klassifikation, K-Means, GMM, etc.) * Untersuchung, wie das Komprimierte Modell die definierte Task auf einem gegebenen Datensatz erfüllt. * Gegenüberstellung KD vs. IR |
| 6 | Schriftliche Ausarbeitung – Vorversion 1-2 | 5 | Bericht (Einleitung, Grundlage, Implementierung, Evaluation, Schluss und Feature Work) | * Version 1 (nach Woche 3) * Version 2 (nach Woche 5) |
|  |  |  |  |  |

**6. Punkt: Methodologie**

Hier sollst Du Deine Arbeitsweise darlegen. Auf welche Methoden beziehst Du Dich und warum? Welchen Begriffsappart wirst Du verwenden oder wird es vielmehr darum gehen, dass Du einen neuen entwickeln wirst? Wie wird sich Deine Recherche gestalten? Wirst Du Experimente oder [Umfragen durchführen](https://gwriters.de/blog/umfrage-erstellen-tools)? Für welche Werkzeuge entscheidest Du Dich? Alle Tätigkeiten, welche für die Umsetzung der Arbeit notwendig sind, sind hier von Relevanz, sollen dargelegt und ihre Anwendung von Dir gerechtfertigt werden.

Zur Lösung der Aufgabenstellung wird die Bearbeitung nach zwei aufeinander Schritten richten. Zuerst wird die Modellkompression auf einen experimentellen Weg untersucht, wo versucht wird, nach Recherchenarbeiten zu bestehenden DNN-Modellkompressionsalgorithmen herauszufinden, wie diese funktionieren, indem sie nachgebaut werden. Diese experimentelle Methode sollte den State-of-the-Art bei dieser Thematik sowie bisher offene Fragen klarstellen.

Nach

**7. Punkt: Planung der Arbeit**

Hier nun wird von Dir bewiesen, dass Du eine Vorstellung hast, wie das Thema angegangen und umgesetzt werden kann.  
Innerhalb welches Zeitraumes werden die einzelnen Schritte stattfinden? Planst Du Puffer ein oder nicht? Zu welchem Zeitpunkt wirst Du voraussichtlich welche Teile der Arbeit fertiggestellt haben? Wer sind Deine Referenzpersonen für Rücksprachen und Aussprachen? Wer sind mögliche Zweitgutachter und in welchem Kommunikationsfluss werden diese mit eingebunden?

Kannst Du diese sieben Punkte bearbeiten, hast Du ein vollständiges Masterabeit-Proposal, das Du einreichen kannst und bist auch inhaltlich bereit, mit dem Schreiben Deiner Masterarbeit zu beginnen. Solltest Du [Unterstützung bei Deiner Masterarbeit](https://gwriters.de/leistungen) brauchen, kannst Du Dich gerne [an unser Team wenden](https://gwriters.de/kontakt).

**06.03.2022**

Hi Christian,

wie besprochen, unser Ziel für deine Masterarbeit:

Knowledge Distillation für effiziente Object Detection auf NVIDIA Jetson Devices.

Mach dich dazu im related work am Besten zu folgenden Themen schlau:

- Was ist State-of-the-Art in Object Detection (zB Yolo V4)? Welche Testdatensätze/Benchmarks gibt es (zB COCO minival)?

- Was gibt es für Object Detection Methoden/Papers auf Edge Devices (Raspberry Pi, NVIDIA Jetson, NVIDIA TX2)

- Was wird für Knowledge Distillation im Bereich Object Detection gemacht? Wo könnten wir vielleicht etwas besser machen? (Ist immer schwierig zu verstehen, aber manchmal wenn man sich frägt wieso ein Paper etwas so macht und nicht anderst)

Wäre klasse, wenn du dazu ein paar Stichpunkte aufschreibst, sodass wir diskutrieren können (zB in Word oder Powerpoint, du musst keine schöne Präsentation bauen)

Wir sprechen nächsten Montag mal über deine ersten Ergebnisse.

Bezüglich der Studienabschlussbeihilfe:

Denke es geht um das Gutachten von einem Professor oder Dozenten. Ich leite es an Prof. Rizk weiter, dieser ist aber bis zum 14. März nicht verfügbar (Urlaub/Geschäftsreise). Alternativ kannst du in der Zwischenzeit auf jeden Fall auch einen anderen Professor anfragen, der dich kennt (zB aus Projekten oder Seminaren, in denen du gut warst). Drücke dir die Daumen!

Sind erwähnte Elemente in Bereichen (Chargiertes Roheinsen und Analysen(Spurenelemente)) spezifisch oder sollen sie gegeben falls auch mehr oder weniger werden. Was ist der Unterscheid zwischen den beiden Bereichen?

Im Schrott-Bereich: Was sind dort max, soll und ist? Wonach werden die Schrotte gefiltert?? (Zeit, Anlage, …)

Object recognition is one of the most important areas and applications of ML and Deep Learning for computer vision. The goal in this area is to identify existing objects in images or videos. This is achieved by agents (ML systems) learning to understand the content of an image. Understanding the content of an image in this context means automatically classifying the image itself in it and locating where present objects in the image appear pixel by pixel. This self-supervised image processing is made in the same way as can be observed in the human occipital lobe, the results of which are commonly used to enable the person to distinguish, for example, between a cat and a dog.

* image classification

Classifying images in the field of intelligent computer vision suggests taking an image as input and outputting a corresponding classification label. The label is based on some important learning metrics like the probability distribution of pixels in the features space built on the entered image, accuracy of the classification model, well as the calculated loss value. Dogs, oranges, and cats can for example on this basis be classified with some probability as elements of classes “animal”, “fruit”.

* Object localization

It is the second task after having classified images. Here, localization algorithms are trying to find out based on pixel distribution being objects in each image. They highlight then each object by using bounding boxes scaled depending on position, height, and width. With bounding box highlighted objects represents the output of a localization algorithm using an image as input.

Combining image classification and object localization takes an image as input generates the same image with calculated classification metrics for muster present in the image. Furthermore, classified musters are bound with the same colored boxes if they look alike and different colored boxes if they are different. Boxes are size-dependent on object size in the input image: That is the task currently archived by object detection algorithms. These algorithms can moreover be effective by multi-class classification and localization.

Therefore, there are some limitations or challenges of using object detection. These occur when considering the rectangular form of the bounding boxes, which is the only form we so far have. Objects containing curvature parts are disadvantaged by shaping objects. Also, some calculations like the determination of perimeters and areas of detected objects are cannot accurately be estimated by object detection algorithms.[[5]](#footnote-5)

One technique for solving the object shaping problem is image segmentation occurring in object detection, another subset of object recognition with more granular bounding contouring than rectangular bounding box-based object recognition. This granularity helps in accurately determining a matched multi-shaped bounding box for detecting objects in an image. Image segmentation is now the state of the art in object detection and has mainly two variants: Instance segmentation and semantic segmentation.

* Instance segmentation

Bounding of objects are identified and pixels located in the same interior area to identified bounding are colored with the same color. Pixels belonging to different bound areas have different coloration.

* Semantic segmentation

Each pixel (including background) in the input image is respectively labeled, based on classes, with different colors.

There are other techniques for object recognition such as template matching or bob analysis, which we will not discuss in the context of this master thesis. The main purpose is to understand how object recognition and more precisely object detection works. For this purpose, we first look for some object recognition algorithms and show how they work.

1. https://link.springer.com/article/10.1007/s11263-021-01453-z [↑](#footnote-ref-1)
2. https://arxiv.org/abs/2101.09671 [↑](#footnote-ref-2)
3. [https://openaccess.thecvf.com/content\_ICCV\_2017/papers/He\_Channel\_Pruning\_for\_ICCV\_2017\_paper.pdf](https://deref-gmx.net/mail/client/gH5paisu_DA/dereferrer/?redirectUrl=https%3A%2F%2Fopenaccess.thecvf.com%2Fcontent_ICCV_2017%2Fpapers%2FHe_Channel_Pruning_for_ICCV_2017_paper.pdf) [↑](#footnote-ref-3)
4. [https://arxiv.org/pdf/1611.06440.pdf](https://deref-gmx.net/mail/client/EuamB5g7YfI/dereferrer/?redirectUrl=https%3A%2F%2Farxiv.org%2Fpdf%2F1611.06440.pdf) [↑](#footnote-ref-4)
5. https://arxiv.org/abs/1311.2524 [↑](#footnote-ref-5)